UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

A Benchmark for Biomedical Knowledge Graph based Similarity

Carlota Maria Alegre Branco Ferreira Cardoso

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Professora Doutora Cátia Luísa Santana Calisto Pesquita

2020

Acknowledgements

First, I would like to thank my supervisor, Professor Cátia Pesquita, for her never-ending support and motivation through the course of this dissertation. Her mentoring was crucial and is one I will never forget. I extend my thanks to Dr. Sebastian Köhler, for his assistance with this project and resultant publications, and Fundação para a Ciência e a Tecnologia for funding this work through the LASIGE Research Unit (UIDB/00408/2020 and UIDP/00408/2020) and the SMILAX project (PTDC/EEI-ESS/4633/2014).

To Diana and Rita, for welcoming me into their academic family, Isabel, for always having her door open for me, and my Biochemistry friends, for still liking me even though I turned to the dark side of Bioinformatics. A special thanks goes to Catarina, for travelling with me through every high and every low: it was a bumpy ride, but we made it.

To everyone in 2º Grupo dos Escoteiros de Portugal, from my cub scouts to my fellow leaders, that have been my second family for the last ten years. I am grateful for every scar, sleepless night, camping trip and endless meeting I have shared with you.

Finally, I would like to thank my mother and brothers for always being there for me, with either a piece of advice or their witty sense of humour. Last, but not least, to my father: I'll make sure to upload this to the cloud.

Resumo

Os grafos de conhecimento biomédicos são cruciais para sustentar aplicações em grandes quantidades de dados nas ciências da vida e saúde. Uma das aplicações mais comuns dos grafos de conhecimento nas ciências da vida é o apoio à comparação de entidades no grafo por meio das suas descrições ontológicas. Estas descrições suportam o cálculo da semelhança semântica entre duas entidades, e encontrar as suas semelhanças e diferenças é uma técnica fundamental para diversas aplicações, desde a previsão de interações proteína-proteína até à descoberta de associações entre doenças e genes, a previsão da localização celular de proteínas, entre outros.
Na última década, houve um esforço considerável no desenvolvimento de medidas de semelhança semântica para grafos de conhecimento biomédico mas, até agora, a investigação nessa área tem-se concentrado na comparação de conjuntos de entidades relativamente pequenos. Dada a diversa gama de aplicações para medidas de semelhança semântica, é essencial apoiar a avaliação em grande escala destas medidas. No entanto, fazê-lo não é trivial, uma vez que não há um padrão-ouro para a semelhança de entidades biológicas. Uma solução possível é comparar estas medidas com outras medidas ou *proxies* de semelhança. As entidades biológicas podem ser comparadas através de diferentes ângulos, por exemplo, a semelhança de sequência e estrutural de duas proteínas ou as vias metabólicas afetadas por duas doenças. Estas medidas estão relacionadas com as características relevantes das entidades, portanto podem ajudar a compreender como é que as abordagens de semelhança semântica capturam a semelhança das entidades.
O objetivo deste trabalho é desenvolver um *benchmark*, composto por *data sets* e métodos de avaliação automatizados. Este *benchmark* deve sustentar a avaliação em grande escala de medidas de semelhança semântica para entidades biológicas, com base na sua correlação com diferentes propriedades das entidades.
Para atingir este objetivo, uma metodologia para o desenvolvimento de *data sets* de referência para semelhança semântica foi desenvolvida e aplicada a dois grafos de conhecimento: proteínas anotadas com a *Gene Ontology* e genes anotados com a *Human Phenotype Ontology*. Este *benchmark* explora *proxies* de semelhança com base na semelhança de sequência, função molecular e interações de proteínas e semelhança de genes baseada em fenótipos, e fornece cálculos de semelhança semântica com medidas representativas do estado da arte, para uma avaliação comparativa. Isto resultou num *benchmark* composto por uma coleção de 21 *data sets* de referência com tamanhos variados, cobrindo quatro espécies e diferentes níveis de anotação das entidades, e técnicas de avaliação ajustadas aos *data sets*.

Palavras Chave: Semelhança semântica, Grafos de conhecimento, *Benchmark*.

Abstract

Biomedical knowledge graphs are crucial to support data intensive applications in the life sciences and healthcare. One of the most common applications of knowledge graphs in the life sciences is to support the comparison of entities in the graph through their ontological descriptions. These descriptions support the calculation of semantic similarity between two entities, and finding their similarities and differences is a cornerstone technique for several applications, ranging from prediction of protein-protein interactions to the discovering of associations between diseases and genes, the prediction of cellular localization of proteins, among others.

In the last decade there has been a considerable effort in developing semantic similarity measures for biomedical knowledge graphs, but the research in this area has so far focused on the comparison of relatively small sets of entities. Given the wide range of applications for semantic similarity measures, it is essential to support the large-scale evaluation of these measures. However, this is not trivial since there is no gold standard for biological entity similarity. One possible solution is to compare these measures to other measures or proxies of similarity. Biological entities can be compared through different lenses, for instance the sequence and structural similarity of two proteins or the metabolic pathways affected by two diseases. These measures relate to relevant characteristics of the underlying entities, so they can help to understand how well semantic similarity approaches capture entity similarity.

The goal of this work is to develop a benchmark for semantic similarity measures, composed of data sets and automated evaluation methods. This benchmark should support the large-scale evaluation of semantic similarity measures for biomedical entities, based on their correlation to different properties of biological entities.

To achieve this goal, a methodology for the development of benchmark data sets for semantic similarity was developed and applied to two knowledge graphs: proteins annotated with the Gene Ontology and genes annotated with the Human Phenotype Ontology. This benchmark explores proxies of similarity calculated based on protein sequence similarity, protein molecular function similarity, protein-protein interactions and phenotype-based gene similarity, and provides semantic similarity computations with state-of-the-art representative measures, for a comparative evaluation of the measures. This resulted in a benchmark made up of a collection of 21 benchmark data sets with varying sizes, covering four different species at different levels of annotation completion and evaluation techniques fitted to the data sets characteristics.

Keywords: Semantic Similarity, Knowledge Graphs, Benchmark.

Resumo Alargado

Atualmente, com o desenvolvimento constante de tecnologias, novas ou pré-existentes, estamos perante um crescimento nunca visto na quantidade de dados produzidos pelos diversos domínios de conhecimento humano. Isto trouxe consigo desafios em lidar com o tamanho, complexidade e diversidade dos dados. Um domínio principalmente responsável por esta explosão de dados, acompanhada pela incapacidade de os processar, foi o domínio das ciências da vida. As diversas tecnologias utilizadas em investigação nas áreas da genómica e proteómica produzem, a uma velocidade superior à de processamento, grandes quantidades de dados complexos sobre a função, regulação e interação de genes e proteínas. O processamento e integração destes dados é fundamental para tarefas como a associação de genes e proteínas às doenças que causam, entre outras aplicações.

O domínio das ciências da vida é composto por áreas diversas que utilizam diferentes e complexas nomenclaturas para expressar o conhecimento que, após processamento dos dados para a sua extração, é guardado sob a forma de linguagem natural. Esta é ainda, aliás, a forma principal de comunicação em ciência, pela divulgação das diversas descobertas feitas no mundo da investigação através da publicação de artigos científicos. No entanto, o crescimento exponencial da quantidade de dados nesta área tornou insustentável a utilização de linguagem natural como a única forma de representação de conhecimento e impulsionou a pesquisa por uma forma estruturada de representação do conhecimento que permitisse a compreensão deste tanto por humanos como por computadores.

As ontologias são um exemplo deste tipo de representações. Estas representam, sob a forma de grafo, os conceitos dentro de um domínio de conhecimento de forma a que cada conceito esteja definido precisamente e através de relações, hierárquicas ou não, com outros conceitos. A capacidade das ontologias de fornecerem uma descrição estruturada das entidades foi o principal motivo que levou à adoção das ontologias para a caracterização de entidades, e à sua proliferação dentro do domínio biomédico e não só.

A caracterização de entidades com ontologias, através do processo de anotação semântica, permite a representação do conhecimento existente sobre estas na forma de um grafo, apelidado de grafo de conhecimento. Os grafos de conhecimento, associados a diferentes técnicas de aprendizagem automática que os explorem, abrem a possibilidade da mineração do conhecimento de domínios complexos, como o biomédico. Ao representarmos várias entidades no mesmo grafo de conhecimento podemos, por exemplo, computar a semelhança semântica entre duas entidades, com algoritmos que explorem e comparem as descrições ontológicas de cada entidade. A semelhança semântica é uma medida da proximidade de significado entre duas entidades, cujo significado é definido pela ontologia que contextualiza o grafo de conhecimento. O cálculo da semelhança semântica suporta a realização de tarefas como integração de dados heterogéneos,

estabelecimento de ligações ou correspondências entre entidades, comparação e agrupamento de entidades ou geração de recomendações. No domínio biomédico, em particular, a semelhança semântica já foi aplicada à previsão de interações proteína-proteína, da associação entre genes e doenças ou de interações fármaco-alvo.

A popularidade destas medidas impulsionou o seu desenvolvimento e, atualmente, existem diversas medidas de semelhança semântica, cada uma com as suas características, e desenvolvidas em diferentes contextos. Esta variedade, tanto em termos de medidas de semelhança semântica como de aplicações para as mesmas, torna fundamental a determinação da melhor medida para cada aplicação. No entanto, esta avaliação não é trivial, uma vez que não existe um padrão-ouro para semelhança entre entidades biológicas e, uma vez que cada medida define a noção de semelhança de forma diferente, determinar qual é a melhor abordagem torna-se uma decisão parcial.

A avaliação de medidas de semelhança semântica e a escolha da melhor medida para cada aplicação é ainda um processo em aberto. Uma vez que a maioria dos estudos realizados neste sentido usam apenas uma medida ou aplicam o estudo a um pequeno, e controlado, conjunto de dados, os seus resultados não são diretamente comparáveis, tornando ainda mais difícil a determinação da melhor medida para a aplicação em questão. Uma vez que não existe nenhuma formal universal de determinar a semelhança entre duas entidades biológicas, uma abordagem possível é o uso de aproximações de semelhança biológica como medida de avaliação das medidas de semelhança semântica. Estas aproximações de semelhança são medidas que capturam a semelhança biológica das entidades através de diferentes ângulos. Por exemplo, ao determinar a semelhança de duas proteínas através da sua sequência, estrutura, função, vias metabólicas e interações estamos a considerar diferentes planos através dos quais um par de proteínas pode ser semelhante ou diferente. Ao utilizarmos estas medidas, estamos a abordar diferentes características das proteínas, que podem ser utilizadas para avaliar como é que cada medida de semelhança captura a semelhança dessas entidades através das diferentes aproximações.

O objetivo desta dissertação é desenvolver um *benchmark*, composto por *data sets* e métodos de avaliação automatizados, que sustentem a avaliação em grande escala e diretamente comparável de medidas de semelhança semântica para entidades biomédicas, com base na sua correlação com diferentes propriedades de entidades biológicas.

Para atingir este objetivo, uma metodologia para o desenvolvimento de conjuntos de *data sets* de referência para semelhança semântica foi desenvolvida. Esta metodologia é composta por quatro passos, nomeadamente seleção de entidades para os *data sets*, formação de pares de entidades para os *data sets*, seleção e computação de medidas de semelhança semântica e biológica entre os pares de entidades e seleção e implementação de técnicas de avaliação das medidas de semelhança semântica. Os primeiros três passos desta metodologia garantem o desenvolvimento de *data sets* de referência que seguem as diretrizes para *data sets* de referência de qualidade e o último passo da metodologia, permite que, com base nos *data sets* desenvolvidos, sejam realizados estudos sistemáticos para a avaliação de medidas de semelhança semântica em grande escala.

A metodologia desenvolvida foi aplicada a dois grafos de conhecimento: proteínas anotadas com a *Gene Ontology* e genes anotados com a *Human Phenotype Ontology*. Este *benchmark* explora aproximações de semelhança calculadas com base na semelhança de sequência de proteínas, semelhança de função molecular de proteínas, interações proteína-proteína e semelhança de genes baseada em fenótipos, e fornece cálculos de semelhança semântica com medidas representativas do estado da arte, para uma avaliação comparativa das medidas. Isto resultou num

benchmark composto por uma coleção de 21 *data sets* de referência com tamanhos variados, cobrindo entidades de quatro espécies diferentes, em diferentes níveis de anotação e incluindo ainda técnicas de avaliação ajustadas às características dos *data sets*. Dos 21 *data sets*, 20 são *data sets* com pares de proteínas anotadas com a *Gene Ontology*, uma vez que esta é a ontologia que é a base de mais medidas de semelhança semântica. Estes *data sets* estão divididos em duas coleções, uma baseada em função molecular e a outra em interações proteína-proteína. Para estes *data sets*, o *benchmark* suporta não só o cálculo da correlação de semelhança semântica com cada uma das aproximações de semelhança disponíveis para cada *data set* como também, para os *data sets* de interação proteína-proteína, avaliação do impacto das medidas de semelhança semântica na previsão destas interações. O *data set* composto por genes anotados com a *Human Phenotype Ontology* contém pares de genes humanos que suportam a avaliação de medidas de semelhança semântica com base na relação destas medidas com a semelhança de fenótipos em que os genes estão envolvidos.

No futuro, o *benchmark* pode ser atualizado através de melhoramentos nos atuais componentes ou introdução de novos. Este tipo de atualizações podem incluir atualização dos *data sets* existentes, através de introdução de novas medidas de semelhança semântica ou biológica, ou novas técnicas de avaliação, expansão dos *data sets* baseados nos grafos de conhecimento utilizados para outro tipo de pares de entidades ou inclusão de novos *data sets* com pares de entidades anotados com outras ontologias biomédicas. Mais ainda, a metodologia desenvolvida nesta dissertação pode ser aplicada a qualquer domínio onde uma aproximação de semelhança possa ser implementada, tornando o desenvolvimento de *benchmarks* análogos fora do domínio biomédico uma possibilidade.

Contents

List of Figures

List of Tables

Acronyms

ABox Assertion Box

BMA Best Match Average

BP Biological Process

CC Cellular Component

CESSM Collaborative Evaluation of Semantic Similarity Measures

ChEBI Chemical Entities of Biological Interest

EC Enzyme Commission

FN False Negative

FP False Positive

GAF Gene Association File

GO Gene Ontology

GOA Gene Ontology Annotation

GP Gene-Phenotypes

HPO Human Phenotype Ontology

IC Information Content

KG Knowledge Graph

MF Molecular Function

OBO Open Biomedical Ontology

OMIM Online Mendelian Inheritance in Man

OWL Web Ontology Language

PCC Pearson Correlation Coefficient

PS Phenotypic Series

PPI Protein-Protein Interaction

RDF Resource Description Framework

SS Semantic Similarity

SSM Semantic Similarity Measure

TBox Terminology Box

TN True Negative

TP True Positive

TSV Tab-separated Values

URI Uniform Resource Identifier

Chapter 1

Introduction

Nowadays, with the non-stop improvement of several technologies and development of new ones, we are witnessing an unprecedented growth in the size of data produced by nearly all domains of human endeavour. That brought with it new challenges in handling the size, complexity and diversity of data. One of the domains where this data deluge has altered nearly every aspect of its workings is the life sciences. High throughput techniques in genomics and proteomics produce large amounts of data about the function, regulation and interaction of genes and proteins, and their integration with clinical research has helped link thousands of genes and proteins to their related diseases, among other tasks (T. P. [2011]).

The biomedical domain is one of complex and volatile data where the knowledge, after its extraction from data analysis, is usually recorded in natural language in scientific publications. However, given the exponential growth of already large quantities of data and associated knowledge, this was an unsustainable form of representation. This knowledge needs to be stored in a computationally amenable fashion, making data and knowledge understandable by both humans and computers (Stevens et al. [2004]).

An example of these types of representations are ontologies. Since the early 2000's, biomedical ontologies have been increasingly used to annotate data, which has resulted in a proliferation of ontologies (there more than 850 currently stored in BioPortal[1] (Whetzel et al. [2011a])) and ontology annotated data sets, many available as linked open data (e.g. Bio2RDF[2](Belleau et al. [2008])). The creation of these resources, as is also the case of more general purpose knowledge graphs (KGs) such as DBpedia (Lehmann et al. [2015]), has been the stepping stone to achieve a structured and linked representation of the knowledge throughout different domains, a web of data.

In the biomedical domain, the ability to compare entities, such as genes, cells, organisms, populations or species, and finding their similarities and differences, is essential to support scientific inquiry. While comparing the sequences of two genes or the structures of two proteins can be achieved directly, because both have objective representations and measurable properties, the comparison of more complex aspects of biological entities, such as their function, is not straightforward.

[1] https://bioportal.bioontology.org/
[2] https://bio2rdf.org/

Ontologies provide mechanisms of objective representation that support measurement of these more complex aspects. Thus, the representation of entities in KGs, coupled with the development of machine learning techniques able to explore them, opens unparalleled opportunities for the mining of complex domains, as the biomedical domain. For instance, having entities represented in the same KG supports the computation of the similarity between entities, with algorithms that explore ontology features, semantic similarity measures (SSMs). These are measures of how closely related in meaning two entities are. Depending on the level of specialization of the ontology used to describe the entities, the spectrum of comparison between the entities will vary. For instance, the Gene Ontology (GO) is an ontology developed for the annotation of gene products, that describes these entities regarding three aspects of gene function: their cellular localization, their function inside the cell and the biological process to which they contribute to (Ashburner et al. [2000]). Two proteins annotated with the GO that are compared with a SSM will undergo a more general comparison than being compared via their structure, sequence or common metabolic pathways.

Several tasks can be supported by these SSMs, such as integration of heterogeneous data, entity linking or matching, comparison and clustering of entities and generation of recommendations (Harispe et al. [2015]). In fact, computing similarity between instances is an integral part of many machine learning techniques, both supervised and unsupervised. In the biomedical domain, semantic similarity (SS) has been successfully applied to such diverse tasks as the prediction of interaction between proteins, of disease-associated genes or of drug-target interaction (Hoehndorf et al. [2015]). It is worth noting that in these applications similarity is not used to detect identity, but rather to predict the likelihood of a given entity exhibiting a given property.

There are several measures available (Harispe et al. [2015]) each with its distinguishing characteristics. Given the variety of approaches and measures for SS, it is fundamental to determine the best measure for each application scenario. However, there is no gold standard for similarity between biomedical entities, and a manual assessment of similarity by domain experts is unfeasible, not only due to the size of the data, but also because each expert is inherently biased towards a viewpoint of the domain or a particular use case. Furthermore, the existing measures formalize the notion of similarity in slightly different ways and for that reason it is not possible to define what the best SSM would be, since it becomes a subjective decision.

Evaluating the reliability of a SSM or determining the best measure for each application scenario is still an active area of research. Most application studies use only one measure, or test them in a small and controlled set of data, developed for that study alone. An unsystematic assessment practice can lead to biases in published results, a phenomenon referred to as the self-assessment trap (Mangul et al. [2019]). Moreover, the results from these studies are not directly comparable across them, making it difficult to assess which measure is best for which purpose.

1.1 Objectives

Because there is no direct way to determine the true functional similarity of two biological entities, one possible solution is to compare the SSMs to other measures or proxies of similarity.

In the biomedical domain, entities can be compared through different lenses. For instance we can compare two genes via their sequence similarity, two proteins via their structural similarity or two diseases by the metabolic pathways they affect. These similarities do not provide the broad spectrum comparison that SS supports, but they can be used as measures of similarity at different levels, since they are known to relate to relevant characteristics of the underlying entities. Thus, correlating SS with such properties can help us understand how well SS approaches capture entity similarity.

Moreover, for an accurate assessment of the impact of a SSM in a given task, the measure should be benchmarked against other SSMs. This means running the same tests for a set of SSMs and comparatively evaluate their performance under the same conditions. This approach would tackle the issues previously exposed in the small scale and non-systematic evaluation of SSMs.

The main goal of this dissertation is to develop a benchmark for SSMs in the biomedical domain. The benchmark should tackle the following challenges:

- Cover multiple KGs;
- Cover entities of multiple species;
- Consider different types of proxy similarities;
- Include representative state-of-the-art SSMs;
- Provide easy to use performance metrics.

1.2 Contributions

The main contributions of this dissertation are:

1. Development of a methodology for the construction of benchmark data sets for KG-based SS;
2. Implementation of the developed methodology to the construction of a collection of benchmark data sets for SS in the biomedical domain. These data sets include pairs of genes or proteins, grouped by species (*D. melanogaster*, *E. coli*, *H. sapiens* and *S. cerevisiae*), annotated with either the GO or the Human Phenotype Ontology (HPO);
3. Development of a benchmark that supports the proxy-based evaluation of SSMs in biomedical KGs;
4. Availability and accessibility of the data sets and benchmark in an online repository[3];
5. Poster paper with the characterization of the benchmark data sets entitled "A Collection of Benchmark data sets for Knowledge Graph based Similarity in the Biomedical Domain", presented at the 17th Extended Semantic Web Conference, online edition, where it was co-awarded best poster paper;

[3] https://github.com/liseda-lab/kgsim-benchmark

6. Submission of an original article entitled "A Collection of Benchmark data sets for Knowledge Graph based Similarity in the Biomedical Domain" to Database: The Journal of Biological Databases and Curation.

1.3 Document Structure

Additionally to the present introductory chapter, that gives a contextualization for the problem at hand and proposed solution, this document is structured in six chapters as follows:

- **Chapter 2** (Concepts) introduces the basic concepts whose understanding is vital for the comprehension of this dissertation, namely, the semantic web, ontologies, KGs, SS and benchmarking.
- **Chapter 3** (Related Work) presents recent work developed in the scope of this dissertation.
- **Chapter 4** (Methodology) presents the general methodology developed for the construction of a benchmark for SS.
- **Chapter 5** (Building the Benchmark) presents the implementation of the methodology from Chapter 4 in two KGs for the construction of the benchmark for SS in the biomedical domain.
- **Chapter 6** (Results and Discussion) presents the results from the methodology implementation and discusses the features of the resulting resource.
- **Chapter 7** (Conclusion) discusses the main conclusions of this work, and indicates some directions for future work.

Chapter 2

Concepts

This chapter introduces the concepts that are vital for the understanding of the work presented in this dissertation. The concepts in here defined are the Semantic Web and its underlying components, SS approaches for ontologies and KGs and guidelines to build quality Semantic Web resources and benchmarks.

2.1 Semantic Web

In Berners-Lee et al. [2001] the Semantic Web is defined as "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation". For it to function, machines must have access to structured and connected collections of data, defined with a strict vocabulary and a set of rules, so that it can be used to conduct automated reasoning by Semantic Web tools.

However, data resulting from biomedical research is stored in word, phrase or text format, and are meaningless for computers that seek to apply reasoning over them. The goal of Semantic Web research is to allow the vast range of web-accessible information and services to be more effectively exploited by both humans and automated tools (Horrocks [2008]). Thus, data should be represented in a formal, structured, machine-readable manner through the process of semantic annotation. This allows meaning to be assigned to the resources, usually by linking them to an ontology, creating a KG (Jacob [2005]; Pulido et al. [2006]).

2.1.1 Ontologies

An ontology is a technique used to represent the knowledge about a domain, by modeling its concepts and the relationships between them (Bodenreider and Stevens [2006]). Ontologies have two components: classes and properties. A class is a term that refers to a set of entities in a system (e.g. the class 'protein' represents all proteins in the system). Properties establish how one class relates to another (e.g. 'protein' 'is a' 'molecule', where 'is a' defines the relationship between 'protein' and 'molecule') (Hoehndorf et al. [2015]). Figure 2.1 shows an excerpt of the GO, a very successful biomedical ontology.

The relations between classes can be structured in triples, as depicted in Figure 2.2. Each class or property in an ontology should be identified by a unique Uniform Resource Identifier

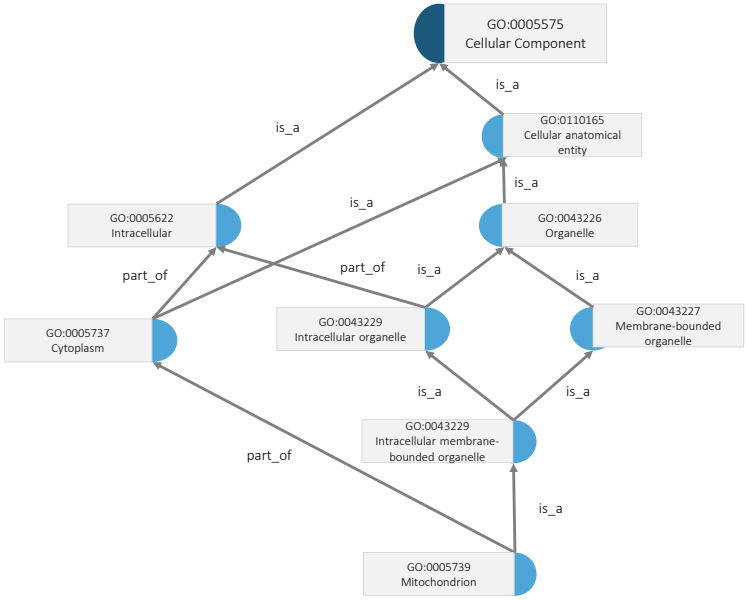


Figure 2.1: Excerpt of the GO graph representing the class GO:0005739 "Mitochondrion" and its ancestors.

(URI) that is used to identify each component of a triple: subject, predicate and object. The predicate denotes the relationship that exists between the subject and the object.

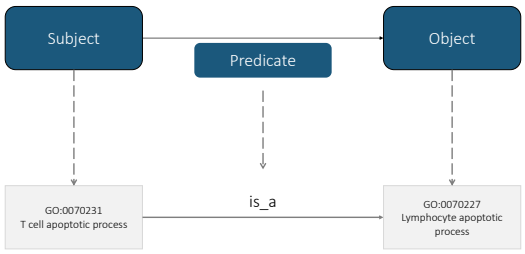


Figure 2.2: Structure and example of a triple from the GO. This triple shows the relation between the class GO:0070321 "T cell apoptotic process" and its ancestor GO:0070227 "Lymphocyte apoptotic process".

These triples make up the backbone of an ontology, and different specialized languages can be used to model them, depending on the requirements and the goals of the applications:

- **Resource Description Framework (RDF):** RDF is a simple language based on triples that specify the relation between the subject and the object via the predicate. A set of RDF triples make up a graph making it possible for RDF to be a valid ontology language, but a limited one (Horrocks [2008]).

- **Web Ontology Language (OWL):** OWL is a vocabulary extension of RDF, but far more powerful. An OWL ontology consists of a set of axioms of 2 types: Terminology Box (TBox) and Assertion Box (ABox). TBox axioms serve the purpose of defining an hierarchy of classes and properties, but also restrictions as the disjointness of two classes or characteristics of some properties. ABox axioms express facts about specific entities (e.g. individuals) and are usually not included in an ontology definition (Horrocks [2008]).

- **Open Biomedical Ontology (OBO) Flat File:** This format attempts to achieve human readability, ease of parsing, extensibility and minimal redundancy. An OBO document's structure is divided in header and stanzas. While the header serves the purpose of describing generic information about the ontology (e.g. version) each stanza enclosures the description and relations of each ontology element (Golbreich et al. [2007]).

In many fields of biomedical research, ontologies play an essential role in tasks such as knowledge representation or data integration. Their acceptance, and relevance of their use in this domain, has been responsible by a non-stopping development of new ontologies, accompanied by the growth of widely used ones in the last years. As of June 2020, Bioportal (Whetzel et al. [2011b]), "the world's most comprehensive repository of biomedical ontologies", stored over 850 biomedical ontologies, with scopes as diverse as the characterization of gene products (as is the GO) to phenotypic abnormalities in human diseases (HPO (Köhler et al. [2009])) or the characterization of drugs (Chemical Entities of Biological Interest (ChEBI))(Hastings et al. [2016])). Ontologies can sometimes have overlapping domains, as is the case of multiple anatomy ontologies, whose goal is the enumeration of existing knowledge about the structure of organisms and their constituting parts. Examples of these ontologies include the Foundational Model of Anatomy (Rosse and Mejino [2004]) for humans and the Drosophila Anatomy Ontology (Mesquita da Costa et al. [2013]) for the fruit fly.

2.1.2 Semantic Annotation

Given the abstract nature of ontologies, their statements are true for all entities of a given type, as opposed to being specific of a certain entity. However, ontologies can be used to describe real-world entities through the process of semantic annotation. Describing entities with classes from an ontology allows the assignment of meaning to these entities. Since ontologies are modelled in a formal and machine readable language, annotating multiple entities with the same ontology allows for reasoning to be applied to them. Figure 2.3 shows part of the graph formed by using the GO to annotate a protein. These annotations can be seen as a semantic description of the protein, since they can be used to, computationally, assign to the protein a meaning.

The process of annotation provides a semantic description of the entities by taking advantage of the hierarchical representation provided by an ontology that enables computational reasoning. This is more than an explicit description of the entity, as one would find in a dictionary entry of a word, as it describes the entities through the classes and their relations to each other. Thus, the richer the ontology is in relations between classes and the thorough the annotation is, the better captured the semantic description of the entity will be (Stevens et al. [2004]).

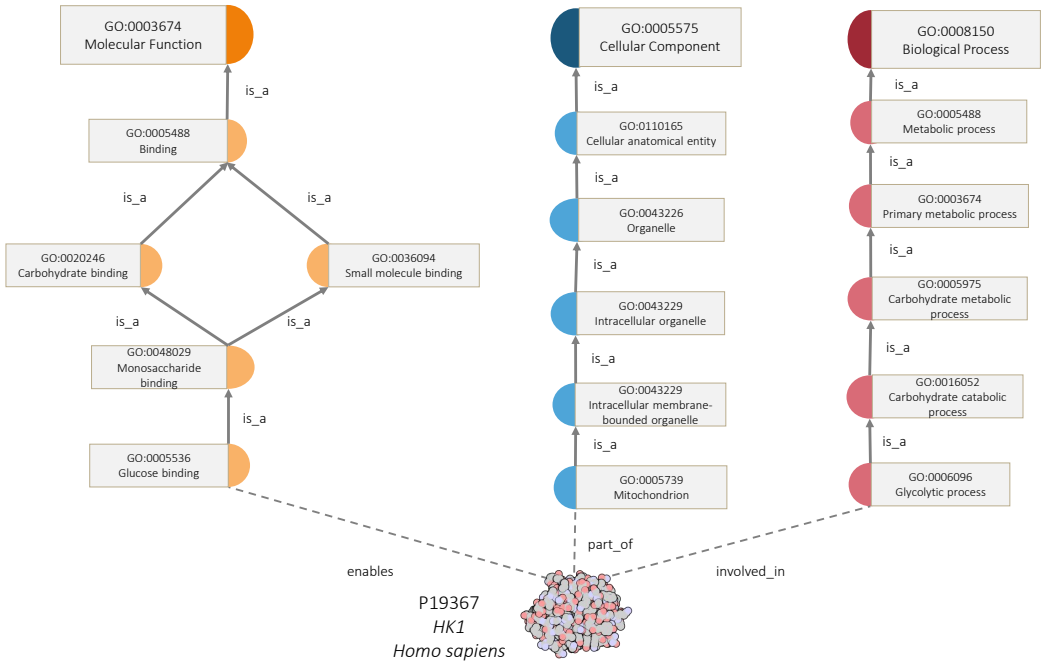


Figure 2.3: Sub-graph of GO formed by a gene product (Protein P19367) and GO classes. For this gene product, the classes that best describe it are chosen for its annotation.

2.1.3 Knowledge Graphs

The knowledge provided by linking entities and ontologies can be represented in graph form. These representations, KGs, are a systematic way to connect the information on entities (e.g. proteins, genes) to data representations. For this purpose, ontologies provide the context in which the entities are being represented (Dörpinghaus and Jacobs [2019]).

The nodes of a KG represent ontology classes or entities while edges represent ontology relations. An example of a portion of a KG is represented in Figure 2.4, contextualized by the GO and its annotations, where proteins are linked to GO classes and other proteins.

KGs have vast applications in the biomedical domain, based on the ontological descriptions of the entities. For instance, enrichment analysis can be performed in a group of interacting proteins to test the hypothesis if a given function is more common in that group of proteins than in a sample of the same size from the whole system (Hunter [2017]).

2.2 Semantic Similarity

A SSM is a function that, given two ontology classes or two sets of classes describing two entities, returns a numerical value reflecting the closeness in meaning between them (Pesquita et al. [2009a]).

The approaches used to quantify SS can be distinguished based on which entities they intend to compare: there are approaches for comparing two classes and approaches for comparing two entities annotated with its own set of classes.

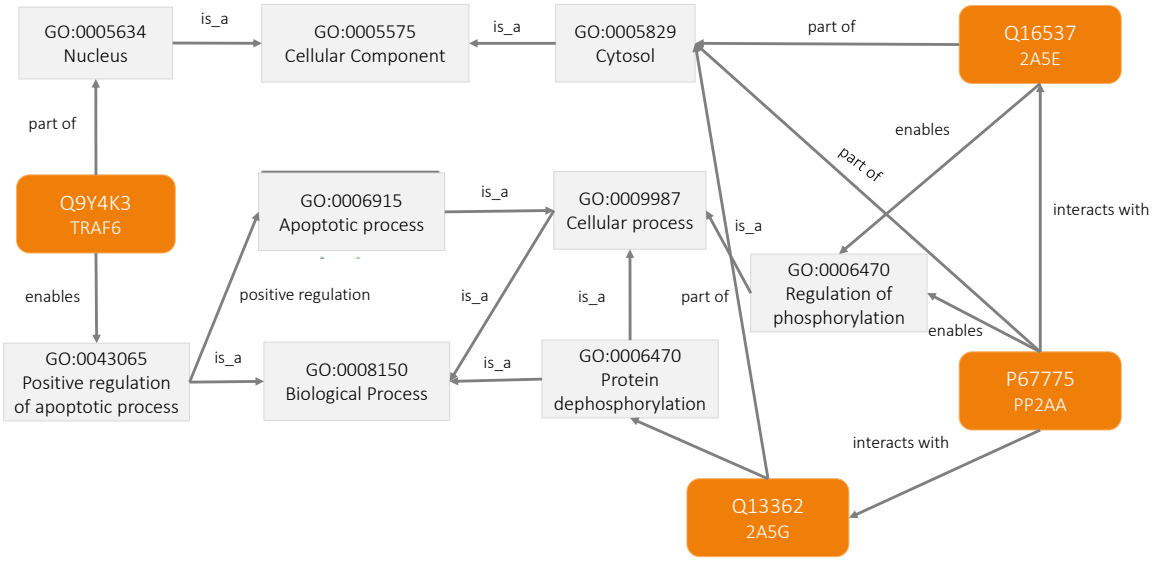


Figure 2.4: Sub-graph of the GO KG illustrating the relations between proteins and GO classes. Proteins are represented in orange boxes and GO classes in grey boxes.

For comparing classes within an ontology, edge or node-based measures can be employed:

- **Edge-based approaches:** These measures rely on the number of edges in the graph path between the two classes, either by calculating the distance between them, or through the length of the lowest common ancestor of the two classes to the root node. Despite being intuitive, these measures bear issues with them: they are based on the assumption that all edges at the same level of the ontology correspond to the same semantic distance between classes, and that edges and nodes are uniformly distributed throughout the graph. These problems make edge-based approaches rarely used in the biomedical domain.

- **Node-based approaches:** These measures depend on the properties of the classes involved. They typically rely on the information content (IC) of a class, a measure of how informative or, rather, specific a class is (Pesquita et al. [2009a]). The IC can be calculated through the graph structure (intrinsic approach) or the number of annotations a class is used on (extrinsic approach).

To calculate SS for two entities, each annotated with a set of classes, both pairwise and groupwise approaches can be used (Pesquita et al. [2009a]):

- **Pairwise approaches:** In these approaches functional similarity between two entities is assessed by combining the SS between their classes.

- **Groupwise approaches:** In these approaches set, vector or graph-based measures are employed.

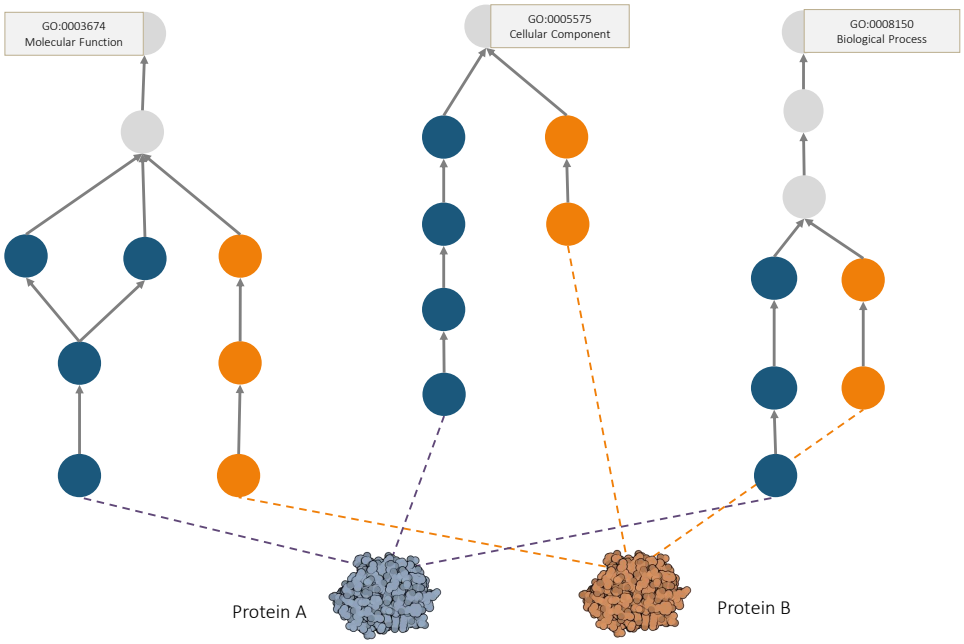


Figure 2.5: Use of the GO annotations for the calculation of SS between two proteins. Grey-filled circles are common classes between the two proteins, blue-filled circles are classes that only annotate Protein A, and orange-filled circles are classes that only annotate Protein B.

From Figure 2.5 we can calculate SS between proteins A and B using a groupwise approach, by finding the number of classes common to both proteins and dividing it by the total number of classes annotating both proteins, i.e., $sim(A, B) = \frac{6}{24} = \frac{1}{4} = 0.25$. This is a very simple tackle on a SS problem, but more sophisticated approaches, combining different strategies for class and entity similarity, can be employed.

SS has been successfully applied to computationally predict protein–protein interactions (PPIs), based on their functional similarity, to the diagnosis of diseases, based on phenotypic similarity, or to the classification of chemicals based on structural similarity (Hoehndorf et al. [2015]). It is worth noting that in these applications, similarity is not used to detect identity, but rather to predict the likelihood of a given entity exhibiting a given property.

2.3 Semantic Web Resources

Scientific advancement relies on quality resources that provide the necessary scaffolding to support scientific publications. Sharing these resources and the best practices that have lead to their development is crucial to consolidate research material, ensure reproducibility of results and, in general, gain new scientific insights.

To integrate the data emerging from all different kinds of research and promote reuse of produced resources, there is a set of guidelines one should follow in their development and publication, known as the FAIR Guiding Principles (Wilkinson et al. [2016]). The FAIR Guiding Principles are a set of guidelines for scientific data management and stewardship that apply to

three types of entities: data (e.g. from gene regulation and expression analysis), metadata (data for the characterization of a digital object, e.g. the GO as a tool for describing gene products) and infrastructure (data platform, e.g. UniprotKB (Consortium [2018])). Given the role of machines in the data to knowledge process, due to the inability of humans to operate at the scale and speed needed by the size and complexity of scientific data, these principles emphasise the importance of machine data management with minimal human intervention and state that these entities should be **F**indable, **A**ccessible, **I**nteroperable and **R**eusable:

- **Findable:** (Meta)data should be easy to find by both users and computers. Data is assigned an unique identifier (e.g. URI), that should also be used when using metadata to describe data (e.g. in the annotation process), to clearly state what metadata is being assigned to each data set and make sure it is described properly.
- **Accessible:** More than found, data should be accessed. This should be ensured either by free or authentication/authorization-based access. Metadata should persist even if data is no longer available.
- **Interoperable:** The data should be described using a formal representation that allows for users to use each other's data. These representations (e.g. ontologies) should also follow the FAIR principles. Additionally, (meta)data should include cross-references to other (meta)data, to create as many meaningful links as possible between them and enrich the knowledge about the data.
- **Reusable:** Ultimately, the goal of the FAIR principles is to allow data reusability. This means ensuring that data can be used by others than its creator. For this purpose, meta(data) should be richly described with relevant attributes that facilitate its reuse (i.e. usage licence, relevant publications, scope, version, etc).

2.3.1 Benchmarking

In bioinformatics, a benchmarking study consists of a comprehensive evaluation of the relative performance of existing algorithms targeted at solving the same, or similar, biology problem (Mangul et al. [2019]).

The wide variety of tools available today, as is the case of SSMs, means that it is often difficult to choose the most appropriate for a specific problem. Comparative evaluation of the different methods is a crucial task, not only to select the most suitable method (e.g. more efficient, more correct, more scalable), but also to evaluate the improvements obtained when a new method is introduced and to identify the strengths and weaknesses of the different algorithms (Aniba et al. [2010]).

Similarly to the FAIR guiding principles for data, there are also a few rules to follow when building trust-worthy benchmarks. These include, but are not limited to, what the benchmark aims at measuring and how it must do so (Aniba et al. [2010]):

- **Relevance:** Benchmarks should be adapted to the algorithms they aim at testing. This means that the evaluating tasks should match the ones the the algorithms are expected

to handle and should capture different abilities of the algorithms. Furthermore, all tasks should produce any type of performance measure that can then be used for comparison of results.

- **Solvability:** The tasks the algorithms are subjected to should be achievable but not trivial. This provides an opportunity for systems to show both their capabilities and their limitations.

- **Accessibility:** The benchmark should be easy to obtain and to use. To promote the comparison of results, these need to be publicly available

- **Independence:** To avoid overfitting, or a bias towards a certain algorithm, the benchmark tasks should not be based on information achieved with the algortihms under evaluation. Instead, independent information from other techniques or from human experts should be used to evaluate the correctness of the results.

- **Evolution:** The benchmark should not stagnate, to prevent researchers from evolving their algorithms solely to achieve a good score in the particular set of tests the benchmark is based on.

These testing and evaluation methods are dependent on data sets, that are used to build the tests the algorithms will undergo. By comparison of the algorithms results to the existing knowledge, their performance can be assessed. For that purpose, benchmark data sets with known and verified outcome are needed.

High-quality benchmark data sets are valuable and may be difficult, laborious and time consuming to generate. From a point of view of reasonable use of resources it is important to share such data sets. Moreover, the reuse of the same data set for different studies, can make them directly comparable, a comparison that would not be reliable otherwise. The criteria for a quality benchmark data set intersect the FAIR guiding principles and criteria for a trust-worthy benchmark study, since these studies rely on the knowledge provided by these data sets, and include the following features (Sarkar et al. [2020]):

- **Relevance**: The data set should include the necessary features for the chosen evaluation task, and avoid the inclusion of non-relevant or indirectly related data.

- **Representativeness**: Representativeness is of special relevance in these data sets. The data set should provide a balanced cross-section of biomedical entities, cover the event space as well as possible and be of sufficient size to allow statistical studies. Positive and negative cases, if applicable, should be included. Additionally, should these data sets be used for supervised learning applications, these predictors will benefit from learning from a more general data set. If the cases used for training are particularly biased towards a feature, the performance of the predictor will be biased as well.

- **Non-redundancy**: This means excluding overlapping cases within each data set.

- **Experimentally verified cases**: Method performance comparisons have to be based on experimental data. This feature is of special importance, once more, in supervised learning applications since, if the data is based on another predictor, the evaluation task would be about the congruence of methods, and not their true performance.
- **Reusability**: The data sets, and related resources, should be shared, to promote their use for other research and publications and their direct comparison.

These set of features will be the guidelines followed when developing and applying the methodology for benchmark construction.

Chapter 3

Related Work

This chapter presents a survey on the work developed in the past years on the scope of this dissertation, namely SS and its applications and tools for the evaluation of KG-based SSMs.

3.1 Semantic Similarity

The use of SS in the biomedical field has benefited many bioinformatics applications by resorting to KG-based SSMs.

GO-based SS is the main focus of investigation of SS in this domain and has been the motivator for the development of several new SSMs. Overall, SS in the GO has been applied mainly for validating and predicting functions and interactions, and for analysing transcriptomics and proteomics data. In predicting and validating the function of gene products, these measures can be combined with other similarity metrics, as structural similarity (Liu et al. [2007]) or sequence similarity (Yu et al. [2016]; Makrodimitris et al. [2018]) for better results. In PPI prediction and validation, similarly, SS can be the sole approach (Jain and Bader [2010]; Zhang et al. [2018]) or be used to improve already existing techniques (Mahdavi and Lin [2007]). Finally, the role of SS in the analysis of transcriptomics and proteomics data is mainly the improvement of clustering of co-expressed gene products (Al-Mubaid and Nagar [2008]; Wang et al. [2005]; Kustra and Zagdanski [2006]).

SS using other ontologies can also play an important role in biomedical research. For instance, the HPO provides comprehensive bioinformatic resources for the analysis of human diseases and phenotypes (Köhler et al. [2018]). Different SS-based techniques can be used to rank diseases annotated with the HPO based on how similar they are to several queries of HPO classes (Köhler et al. [2009]; Gong et al. [2018]). This method allows for the clinical diagnosis of patients, by finding the most similar disease to their set of symptoms. Similarly, Masino et al. [2014] use SS and the HPO to predict the disease causing gene in patients. Some SS-based approaches in the HPO resulted in the development of SSMs (Köhler et al. [2009]; Xue et al. [2019]; Hoehndorf et al. [2011]).

Another example of an ontology with use in the biomedical domain is ChEBI (Hastings et al. [2016]). ChEBI is a database and an ontology, that contains information about chemical entites, where these are classified, within the ontology context, based on their structure and biological

or chemical role. Despite being scarcer, there are some applications for ChEBI-based SS. For instance, combining semantic and structural similarity can be used to predict molecular function (Ferreira and Couto [2010]) or drug repurposing (Tan et al. [2014]).

The popularity and diverse utility of SS led to the development of several SSMs. Tables 3.1 and 3.2, an updated adaptation of the survey in Guzzi et al. [2011], present an overview of different SSMs, with its original publication and important features.

Table 3.1: Summary of Pairwise SSMs. Columns Class IC, Common Ancestor (MICA), All Common Ancestors (ACA), Path Length and Class Depth refer to the different features of SSMs. Column Evaluation Technique describes the context in which each similarity measure was tested.

Name	Reference	Class IC	MICA	ACA	Path Lenght	Class Depth	Evaluation Technique
Annotation cosine	Bodenreider et al. [2004]	No	No	No	No	No	Associations between classes
BSM	Cheol Jeong and Chen [2015]	No	Yes	Yes	No	Yes	Correlation with sequence; Functional clustering
EISI	Zhang and Lai [2014]	Yes	No	No	No	No	Intra-pathway similarity
G-SESAME	Wang et al. [2007]	No	No	Yes	Yes	No	Gene Clustering
GraSM	Couto et al. [2007]	Yes	Yes	No	No	No	Correlation with MF
GOGO	Zhao and Wang [2018]	Yes	No	Yes	No	Yes	Gene Clustering
HRSS	Wu et al. [2013]	Yes	Yes	No	Yes	Yes	Correlation with EC, MF and sequence
Jiang and Conarth	Jiang and Conrath [1997]	Yes	Yes	No	No	No	Correlation with human perception
Lin	Lin [1998]	Yes	Yes	No	No	No	Correlation with MF
Othman	Othman et al. [2008]	Yes	Yes	No	Yes	Yes	Annotation prediction
PS or PK-TS	Pekar and Staab [2002]	No	Yes	No	Yes	Yes	Classification of synonyms
Resnik	Resnik [1995]	Yes	Yes	No	No	No	Correlation with human perception
RSS	Wu et al. [2005]	No	Yes	No	Yes	Yes	MF prediction
SB-TS	Yu et al. [2005]	No	No	No	No	Yes	MF prediction
SimDEF	Pesaranghader et al. [2015]	No	No	No	No	No	Correlation with sequence and gene expression
SimIC	Li et al. [2010]	Yes	Yes	No	No	No	PPI prediction
SPBHM	Bandyopadhyay and Mallick [2013]	No	No	Yes	Yes	No	PPI prediction; Correlation with gene expression
simRel	Schlicker et al. [2006]	Yes	Yes	No	No	No	MF prediction
SSDD	Xu et al. [2013]	No	No	Yes	Yes	No	Correlation with EC, MF and sequence
SSM	Couto et al. [2003]	Yes	Yes	No	Yes	Yes	Correlation with protein structure
TCSS	Jain and Bader [2010]	Yes	Yes	No	No	No	PPI prediction
TopoICSim	Ehsani and Drablos [2016]	Yes	No	Yes	Yes	No	Correlation with EC, MF and sequence
Wu	Wu et al. [2005]	No	No	Yes	No	No	MF prediction
Wu-Palmer	Wu and Palmer [1994]	No	Yes	No	Yes	Yes	Translation selection
XOA	Sanfilippo et al. [2007]	Depends on measure used					Correlation with sequence

3.2 Evaluation of Semantic Similarity Measures

Given the variety of approaches and measures for SS, it becomes fundamental to evaluate how well each measure captures the true similarity between two entities. However, the wide range of domains and types of entities SS is applied to, makes it non-trivial to do so.

Based on the few comparative studies that exist, Pesquita [2017] identified the most successful measures in different applications of GO-based SS (Table 3.3). However, in these studies, only a few measures of interest for these particular applications are tested.

Although more classic measures of SS such as Resnik still provide top results in some settings, there is a newer generation of measures that provide better results. These are a part of the new

Table 3.2: Summary of Groupwise SSMs. Columns Class IC, Common Ancestor (MICA), All Common Ancestors (ACA), Path Length, and Class Depth refer to the different features of SSMs. Column Evaluation Technique describes the context in which each similarity measure was tested.

Name	Reference	Class IC	MICA	ACA	Path Lenght	Class Depth	Evaluation Technique
Aggregative approach	Ferreira and Couto [2019]	Depends on measure used					Annotation prediction
Ali and Deane	Ali and Deane [2009]	No	Yes	No	No	No	Function prediction
Cho	Cho et al. [2007]	Yes	Yes	No	No	No	PPI prediction
Cosine	Popescu et al. [2006]	No	No	No	No	No	Correlation with sequence
Czekanowski-Dice	Martin et al. [2004]	No	No	Yes	No	No	Gene Clustering
Dice	Popescu et al. [2006]	No	No	Yes	No	No	Correlation with sequence
FMS	Wu et al. [2005]	Yes	No	No	No	No	MF Prediction
ICOR	Chen et al. [2012]	Yes	No	Yes	No	No	Correlation with EC, MF and sequence
Integrative approach	Ferreira and Couto [2019]	Depends on measure used					Annotation prediction
IntelliGO	Benabderrahmane et al. [2010]	Yes	Yes	No	Yes	Yes	Inter-set cohesion
Jaccard	Popescu et al. [2006]	No	No	Yes	No	No	Correlation with sequence
Kappa statistics	Huang et al. [2007]	No	No	Yes	No	No	Gene functional enrichment
NTO	Mistry and Pavlidis [2008]	No	No	Yes	No	No	Correlation with sequence
PL	Al-Mubaid and Nagar [2008]	No	No	No	Yes	No	Gene Clustering
simGIC	Pesquita et al. [2008]	Yes	No	Yes	No	No	Correlation with sequence
simLP	Gentleman et al. [2004]	No	Yes	No	No	Yes	Gene distance
simNLP	Ye et al. [2005]	No	Yes	No	No	Yes	MF prediction
simUI	Gentleman et al. [2004]	No	No	Yes	No	No	Gene Distance
SSA	Sheehan et al. [2008]	Yes	Yes	Depends on measure used			Gene Clustering
SORA	Teng et al. [2013]	Yes	No	Yes	No	Yes	Correlation with EC, MF and sequence
TO	Lee et al. [2004]	No	No	Yes	No	No	Gene Clustering
TAS	Yu et al. [2007]	No	Yes	No	No	No	Gene functional enrichment
Weighted cosine	Chabalier et al. [2007]	Yes	No	No	No	No	Prediction of functional networks
WJ	Popescu et al. [2006]	Yes	No	Yes	No	No	Correlation with sequence

Table 3.3: Most successful measures in different applications of GO-based SS.

Standard	Best Measure
Sequence similarity	SSDD, SimGIC, HRSS
Pfam similarity	SORA, SSDD, SimGIC
EC similarity	SSDD, HRSS, SORA
Expression similarity	TCSS, SimGIC, SimIC, BMA (Resnik)
Protein–protein interaction	TCSS, SimIC, Max (Resnik)

wave of more complex structural-based measures, such as SSDD, SORA and TCSS, that are now on the lead.

Some semantic web related applications have turned to crowd-sourcing (e.g., in ontology matching (Cheatham and Hitzler [2014]) and in verification of relations (Mortensen et al. [2013])), which brings with it a series of new challenges. The evaluation task is highly dependent on the ability to provide crowd-sourced workers with enough information to make a decision and can be inherently biased towards a particular viewpoint of the domain or a particular use case. This is of extreme relevance in biomedical SS, where two entities may be deemed similar across

many different axis. For instance, two proteins can be compared via their molecular function, expression, cellular localization, disease involvement, etc., making it difficult to select the best angle of evaluation.

Most SSMs evaluations studies are data set based, where a new SSM is applied to all entities or pairs in that data set and its performance (e.g. its correlation to some property, its F-measure when used in a classification task or its p-value when used for the testing of an hypothesis) is assessed, usually by comparison with those of other SSMs.

Building a gold standard data set to support SS evaluation is not trivial. Accomplishing this manually is extremely time consuming and existing manual gold standards are very small compared to the size of the ontologies they correspond to. For instance, Pedersen et al. [2007] created a set of only 30 term pairs extracted from Unified Medical Language System (UMLS) Metathesaurus. The UMLS contains over one million biomedical concepts and five million concept names, originating from more than 100 incorporated controlled vocabularies and classification systems.

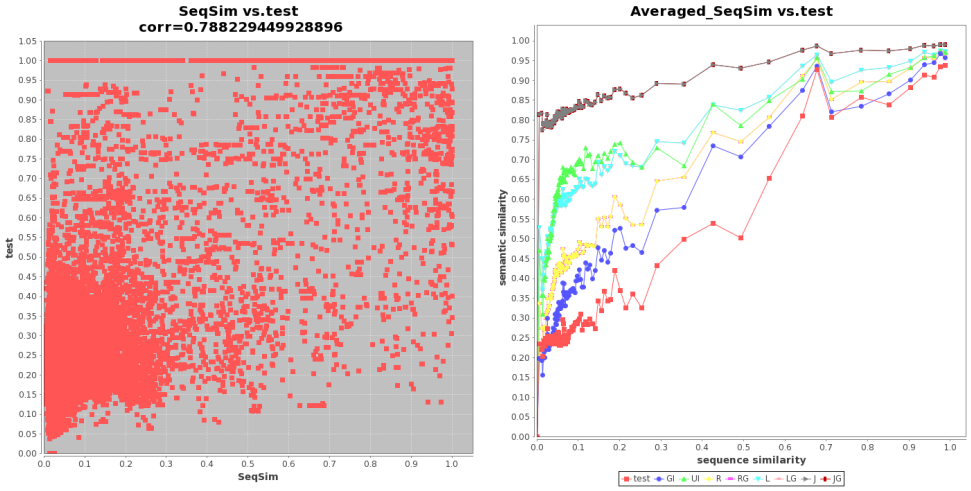
Many data sets have been used to evaluate KG-based SS. MateTee (Morales et al. [2017]) was evaluated both with data set of proteins annotated with the GO and with a data set based on DBpedia entities of the type Person. AnnSim (Palma et al. [2015]) was also evaluated with the same protein data set, but it also included additional evaluations based on disease similarity and drug-target interaction prediction using other data sets.

Most of these studies are independent, meaning their results are not directly comparable, and since they do not provide the necessary KG annotations for each data item, that are likely to change over time, a fair and unbiased comparison between different tools would need to be run with the exact versions of the data.

Pesquita et al. [2009b] developed CESSM (Collaborative Evaluation of Semantic Similarity Measures). CESSM is a webtool designed for the comparison and evaluation of new GO-based SS measures against previously published ones, considering their relation to sequence, Pfam (El-Gebali et al. [2018]) and Enzyme Commission (EC) (Bairoch [2000]) similarity. CESSM's data set is made up of pairs of well-characterized proteins, regrading their classes average IC and existence of Pfam families and EC classification.

Figure 3.1 provides an example of the evaluations provided by CESSM: Pearson correlation between the novel measure and a molecular function similarity proxy - sequence similarity (Figure 3.1a) and the comparison of the behavior of the novel and state- of-the-art measures against sequence similarity (Figure 3.1b).

CESSM was released in 2009, updated in 2014 and since then been widely used by the community, being adopted to evaluate over 25 novel SSMs developed through different methods, more recently common IC-based metrics (Paul and Anand [2018]), and based on vector representations/graph embeddings (Zhong et al. [2019]). CESSM was built as web-based tool to support the automatic comparison against the benchmark data. Over time some limitations of its use were identified: users looking to perform iterative evaluations were limited by access through a graphical user interface; users were unable to calculate other metrics of performance not supported by the tool; users were limited to a single ontology (GO) and a single functional perspective given by the Pfam and EC proxies which focus on molecular function similarity.



(a) Sequence similarity plot against the novel measure.

(b) Averaged sequence similarity plot against several SSMs.

Figure 3.1: Example of CESSM plot results

Despite its limitations, CESSM's methodology for evaluation of SSMs shows effectiveness, since all works developed with the same data are directly comparable, and its data set can be used for other tasks than the evaluation techniques of SSMs supported by the tool (Liu et al. [2018]).

Although not directly related to biomedical SS, there are related contributions in the area of benchmark data for link prediction (Bordes et al. [2013]; Socher et al. [2013]) and classification in knowledge graphs (Ristoski et al. [2016]). KG-based SS can be applied in these contexts, but these benchmark data sets do not support a direct evaluation of SSMs.

Chapter 4

Methodology

This chapter presents an overview of the developed methodology for the creation of the benchmark and its use in the evaluation of SSMs.

The developed benchmark is composed of data sets and automated SS evaluation methods. The first step, and the most important one, is defining the scope of each data set and what measures of biological similarity will be used for the evaluation of SSMs. Only then can the benchmark data sets and evaluation techniques be built around them. In this dissertation, a general methodology for the construction of benchmark was developed. This methodology follows four basic steps:

1. Definition of criteria for entity selection;
2. Definition of criteria for pair formation;
3. Computation of similarity measures for each pair of entities;
4. Development of automatic similarity measures evaluation techniques.

Figure 4.1 represents an overview of the basic methodology for the development of the benchmark, with special focus on the development of the data sets.

Following the development of the data sets, automated evaluation methods should be built to test the performance of the SSM. These methods should be based on the properties of the data sets.

4.1 Definition of Criteria for Entity Selection

When selecting entities from the KG to the benchmark data sets, one should base the decision on two features: the existence of of the necessary information to support the calculation of proxy similarities, and the completion of the annotation of the entities.

Since these data sets will support proxy-based evaluation of SSMs, the first, and most important, criteria for the selection of entities is the existence of the necessary information for the calculation of these proxies. Without that information, a proxy-based evaluation of the SSM calculation would be impossible. For instance, let us imagine two orthologous proteins, A and B. These two proteins are inferred to be descended from the same ancestral sequence separated

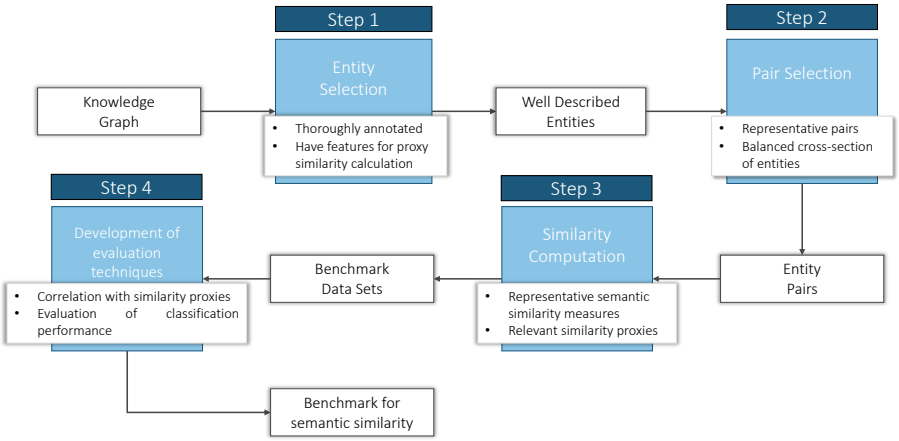


Figure 4.1: Overview of the methodology steps for the development of benchmark.

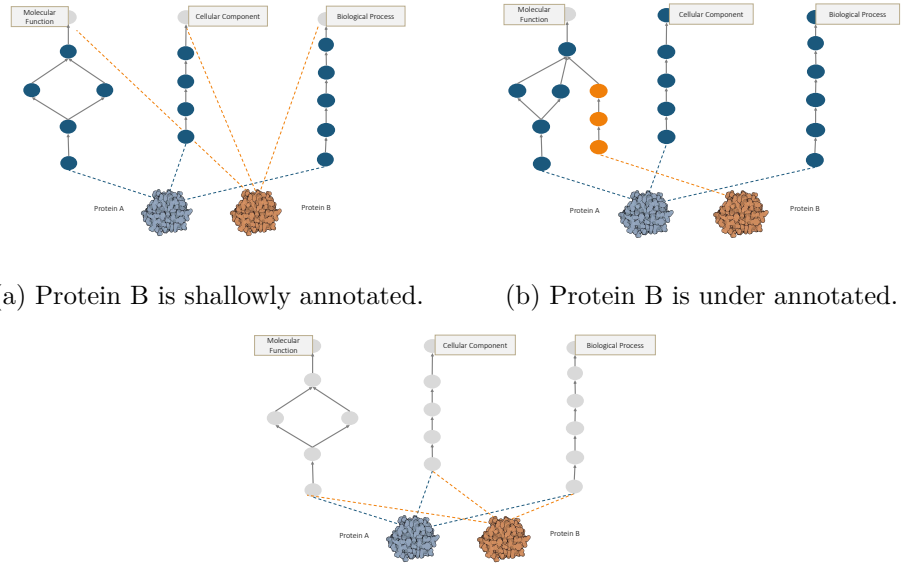
by a speciation event, thus they share high sequence and function similarity. However, if the sequence information for one of these proteins is not available, we can not calculate sequence similarity and assess if the SSM value is consistent with the other representations of similarity.

Additionally, the annotations of these entities, namely their number and the ontology level they are placed on, are also relevant for the SSM performance. Shallow annotations will influence the SS results for the pairs of entities (Pesquita [2017]). Ignoring the specialization level of the classes in the annotation can result in SS results that are inconsistent with human perception. Thus, entities should be well described with classes from the ontology that contextualizes the KG they are represented in, in order to avoid unrealistic SS values.

On the other hand, the number of annotations an entity carries can also impact SS calculation. For instance, an entity described with only one class is most likely to under perform in pairwise SS approaches, as only that class will be compared against all the classes describing the other entity.

Figure 4.2 compares the impact of different levels of ontology description in SS. Let us imagine the same two analogous proteins, A and B. We know that due to their common ancestry, they share the same functions. However, a poor annotation description will result in unrealistic SS values, when compared to the other similarity representations (e.g. sequence similarity). In panel 4.2a Protein B is shallowly annotated, SS is almost null. In panel 4.2b Protein B is under annotated, SS is low. Finally, in panel 4.2c Protein B is properly described and the SS value is accurate.

In sum, candidate entities for the benchmark data sets should have their meaning well captured through ontology annotations. Other levels of annotation completion could also be explored. These would provide a new axis of evaluation of SSMs, that of evaluating the impact of the level of annotation completion in the ability of a measure to capture similarity between entities. However, this is out of scope for this dissertation and will not be considered.



(a) Protein B is shallowly annotated.

(b) Protein B is under annotated.

(c) Protein B is properly described.

Figure 4.2: Different levels of annotation descriptions and its impact on SS calculation. White-filled circles are common classes between the two proteins, blue-filled circles are classes that only annotate Protein A, and orange-filled circles are classes that only annotate Protein B.

4.2 Definition of Criteria for Pair Formation

The selected entities are then combined to form pairs. These pairs will make up the benchmark data sets and be the basis for the calculation of SS. Thus, their selection can not be done randomly. Entity pairs, as a whole, must provide a cross-section of the pairs of entities of each species, while being representative of similarity values and excluding overlapping cases.

4.3 Computation of Similarity

In the last step of the development of the benchmark data sets SS and proxy similarity measures are calculated for each pair generated in the previous step. The SSMs used should be representative of different types of approaches and well accepted in the community, while the similarity proxies should be based on the properties of the entities so that they can, ultimately, be used for the evaluation of SSMs. These proxy measures should be selected attending to the selected KG characteristics, i.e. if the entities are diseases, proxy similarity cannot be sequence similarity but rather involved genes.

These first three steps, ensure the development of data sets that respect the criteria for quality benchmark data sets, namely relevance, representativeness and non-redundancy, through the constraints in the selection of entities, pairs and similarity measures.

4.4 Automated Evaluation Methods

In the last step of the construction of the benchmark, automated evaluation methods are developed. These methods should evaluate SSMs through their relation to the similarity proxies selected in the development of the benchmark data sets.

The steps for the evaluation of SSMs using the benchmark are as follows:

1. Selection of SSM and benchmark data sets;
2. Calculating the selected SSM for all the pairs in the chosen data sets;
3. Evaluating the SSM performance, by performing the same testes for the new measure and the state-of-the-art SSMs.

To do so, different evaluation methods can be developed, based on the type of similarity proxies the data set holds. Figure 4.3 shows examples of evaluation methods the SSM can undergo based on different types of similarity proxies.

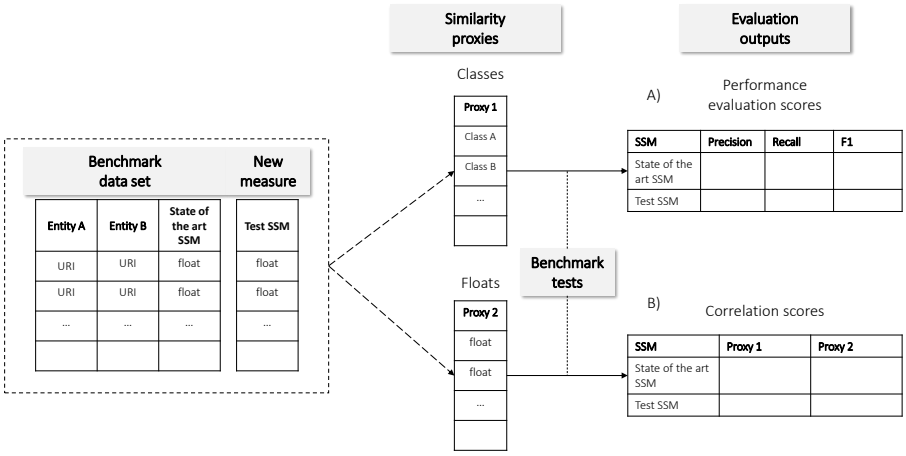


Figure 4.3: Examples of automated evaluation methods supported by the benchmark data sets. A) SS-based classification evaluation B) Proxy-based correlation calculation.

The SSM under evaluation and the state-of-the-art SSMs available in the benchmark data sets, must always be subjected to the same tests, so that their performance can be directly comparable and any relevant findings made publicly available.

Chapter 5

Building the Benchmark

This chapter describes the application of the methodology presented in Chapter 4 to the creation of a benchmark for SSMs in the biomedical domain, based on benchmark data sets. For the construction of the data sets, two KGs were selected: proteins annotated with the GO and human genes annotated with the HPO. Three sets of benchmark data sets resulted from this: PPI data sets, molecular function (MF) data sets and gene-phenotypes (GP) data sets. PPI and MF data sets are made up of protein pairs, while the GP data set is composed of gene pairs. Adequate evaluation methods for each group of data sets were also developed under the scope of the creation of this benchmark. The development of this resource followed the FAIR Guiding Principles for scientific data management and stewardship.

5.1 Gene Ontology-based Benchmark data sets

5.1.1 The Gene Ontology Knowledge Graph

The GO is the most successful case of the use of an ontology in biomedical research and it is used for the annotation of gene products. GO is a directed acyclic graph that covers three distinct aspects of gene function: MF, Cellular Component (CC) and Biological Process (BP). The MF sub-ontology describes the gene product's role at the molecular level, CC describes the cellular location of a gene product's activity and BP the larger biological program to which a gene's MF contributes (Ashburner et al. [2000]). The majority of gene products are proteins, and here on after will refer solely to proteins as the instances described by the GO.

The Gene Ontology Annotation (GOA) is a project that aims to provide assignments of GO classes to proteins of any species, i.e., the semantic annotation of the proteins (Huntley et al. [2014]). The annotations provide a standardized way to describe the regular activity of proteins, and make them directly comparable with SSMs. The classes selected to annotate a protein must always be the most specific classes in the ontology that best describe the entity's function according to available evidence because, due to the hierarchical structure of the GO, a gene product that is annotated with any GO class is also annotated with all its ancestors. Moreover, annotations are are a snapshot of the knowledge about an entity at a certain point in time, meaning they can change over time, to reflect changes in knowledge and/or changes in the ontology.

The integration of GO with the GOA graph results in a KG where the instances are proteins that are linked to GO via their annotations. Figure 2.3 shows an example of the use of GO to annotate a gene product, the protein P19367. Each annotation is supported by an GO Evidence Codes from the Evidence and Conclusions Ontology and a reference.

5.1.2 Proxies of Protein Similarity

For the protein data sets, three proxies of protein similarity: sequence similarity, MF similarity, and PPI. These measures capture different aspects of protein similarity and are typically used for the evaluation of SSMs, usually by correlation assessment (in the case of MF and sequence similarity) or evaluation of the predictive power of SS (in the case of PPI) (Pesquita [2017]).

Protein sequence similarity measures the relationship between two sequences and it establishes the likelihood for sequence homology. Sequence similarity (sim_{seq}) was calculated through the relative reciprocal BLAST score (RRBS) (Pesquita et al. [2008]), given by

$$RRBS(A,B) = \frac{BLAST_{bitscore}(A,B) + BLAST_{bitscore}(B,A)}{BLAST_{bitscore}(A,A) + BLAST_{bitscore}(B,B)} \tag{5.1}$$

where A and B are two proteins.

The relationship between sequence similarity and SS is non-linear (Pesquita et al. [2008]) but becomes more relevant the higher the sequence similarity is (Ikram et al. [2018]). While sequence similarity can be used to evaluate the performance of a SSM, it should not be the sole evaluator in this task.

MF similarity is computed by comparing the functional regions (commonly termed domains) that exist in each protein sequence. Protein functional domains are extracted from the Pfam references contained in the UniProt database (El-Gebali et al. [2018]). Pfam similarity (sim_{Pfam}) is calculated as a Jaccard similarity, using the ratio between the number of families common to proteins A and B and the total number of distinct families through proteins A and B, with

$$sim_{Pfam}(A,B) = \frac{|f_A \cap f_B|}{|f_A \cup f_B|} \tag{5.2}$$

where A and B are two proteins with the set of families f_A and f_B.

The more functional domains two proteins share, the more likely will be that their SS in the GO is high, especially in the MF aspect since these domains are usually responsible by assigning functions to proteins.

PPI has a binary representation: 1 if the proteins are reported to interact, 0 otherwise. Two proteins are considered to be similar if they interact. PPIs have some correlation to SS in the GO: if two proteins are co-localized in the cell and involved in the same large scale process, they are most likely to interact and will share some GO classes in the BP and CC aspects. Both Sousa et al. [2020] and Maetschke et al. [2011] show this to be true with PPI predicting approaches that demonstrate higher predictive power when using classes from these two aspects. However, two proteins can be very similar through different lenses (for instance, having high sequence, semantic or MF similarity), due to being orthologous proteins, but not interact.

The above mentioned similarity proxies were employed after the generation of pairs for the data sets, according to the scope of each collection of data sets. Data sets in the MF collection

employ only MF and sequence similarity, while data sets in the PPI collection employ PPI and sequence similarity.

5.1.3 Selection of Pairs of Proteins for the Benchmark data sets

The protein benchmark data sets are constituted by pairs of proteins, each identified by their UniprotKB Accession Number (Consortium [2018]), and annotated with classes from the GO. The GO graph (dated October 2019) was collected from the website[1] in OBO format and contains 47,413 ontology classes. GO annotations were downloaded from the GOA[2] database for four species in Gene Association File (GAF) 2.1 format. The selected species were *D. melanogaster*, *E. coli*, *H. sapiens* and *S. cerevisiae* due to being highly studied organisms with a large number of annotations in each species GOA graphs. All GOA graphs are dated September 2019, except *E. coli* dated July 2019. The species GOA graphs were made up of 12,490 *D. melanogaster*, 5,341 *E. coli*, 19,464 *H. sapiens* and 6,048 *S. cerevisiae* proteins, respectively, and their GO annotations. Information for the calculation of the similarity proxies, namely Pfam families and protein sequence was retrieved from each protein's UniprotKB[3] entry.

In GO SS, not only is the depth of the GO classes an important feature, as previously exposed, as is the breadth of annotations within the three GO aspects. This is relevant not only because measures may wish to handle the aspects differently, but also because each sub-ontology describes an important functional aspect of the semantic characterization of a gene product, and its characterization is only complete when including classes from the three aspects. For instance, two proteins sharing a specific CC, will share a GO class with high IC. However, being co-localized is not feature enough to determine if these two proteins are, in fact, similar, because they can play different roles in that compartment and in the organism as a whole. This analysis is parallel to the other GO aspects: two proteins can perform the same MF in different cellular compartments and contribute with that function to different biological processes or they can be a part of the same BP and not contribute with the same function or perform it in the same cellular structure. If only one semantic aspect of the gene products is captured, the characterization of the entity will be incomplete and can lead to a misleading similarity assessment between these two proteins. Thus, full coverage of the GO aspects is crucial in choosing the entities.

Given the importance of both depth and breadth in the annotations, the proteins of each of the four selected species were filtered according to the following criteria:

- *One aspect*: The proteins must have at least one annotation in each GO aspect, and in at least one aspect there should be at least one leaf-class annotation.
- *All aspects*: The proteins must have at least one annotation in each GO aspect, and in each aspect there should be at least one leaf-class annotation.

This ensures that all proteins are sufficiently annotated to support SS calculations in either one or all aspects of GO. These criteria also result in all proteins in the *All aspects* data set being included in the *One aspect* as well.

[1] http://geneontology.org/
[2] https://www.ebi.ac.uk/GOA/downloads
[3] https://www.uniprot.org/

For the creation of the PPI data sets, a set of well-known benchmark PPI data sets were used to extract interacting and non-interacting pairs. These data sets comprise pairs of proteins with information about their interaction. Their original publication references and species are presented in table 5.1.

Table 5.1: Benchmark PPI data sets used to select protein pairs for the PPI data sets with original publication reference and protein's species.

Data set	Original Publication	Species
STRING-SC	Maetschke et al. [2011]	*S. cerevisiae*
STRING-HS	Maetschke et al. [2011]	*H. sapiens*
STRING-EC	Maetschke et al. [2011]	*E. coli*
STRING-DM	Maetschke et al. [2011]	*D. melanogaster*
DIP-HS	Jain and Bader [2010]	*H. sapiens*
BIND-SC	Ben-Hur and Noble [2005]	*S. cerevisiae*
DIP/MIPS-SC	Ben-Hur and Noble [2005]	*S. cerevisiae*
GRID/HPRD-bal-HS	Yu et al. [2010]	*H. sapiens*
GRID/HPRD-unbal-HS	Yu et al. [2010]	*H. sapiens*

The pairs in which both proteins met the required criteria were then grouped by species (*D. melanogaster*, *E. coli*, *H. sapiens* and *S. cerevisiae*), excluding all existing duplicate pairs. Two additional data sets, joining all the pairs from these data sets at the same level of annotation completion, were created. For all the pairs in these data sets, the corresponding similarity proxies (PPI and sequence similarity) were calculated.

For the creation of the MF benchmark data sets, species-specific data sets (*D. melanogaster*, *E. coli*, *H. sapiens* and *S. cerevisiae*) were created. Since MF similarity is based on Pfam assignments to proteins, proteins in these data sets should also have, at least, one domain identified in the Pfam database, which further filters down the number of eligible proteins.

Pairs of proteins were randomly generated ensuring roughly the same number of pairs with null and total Pfam family identity. Two additional data sets, joining all the pairs from the these data sets at the same level of annotation completion, were created. For all the pairs in these data sets, the corresponding similarity proxies (MF and sequence similarity) were calculated.

5.2 Human Phenotype Ontology-based benchmark data set

5.2.1 Human Phenotype Ontology Knowledge Graph

The HPO provides comprehensive bioinformatic resources for the analysis of human diseases and phenotypes (Köhler et al. [2018]). It is organized as independent subontologies that cover different categories: "Phenotypic Abnormality" contains descriptions of clinical abnormalities, "Mode of Inheritance", "Frequency" and "Clinical Course" describe the relation between patients

or diseases and their symptoms, and "Clinical Modifier" is designed to characterize and specify the phenotypic abnormalities defined in the "Phenotypic Abnormality" subontology (Köhler et al. [2018]). The HPO has been used to integrate sequencing data from multiple biotech centres to identify patients with mutations in the same gene and comparable phenotypes, to record detailed clinical phenotypes of patients with rare inherited disorders, to annotate clinical cases with standard phenotype variants in order to cluster phenotypically overlapping patients, and finally to increase interoperability between clinical laboratories (Köhler et al. [2018]). The HPO can be used to annotate both patients, diseases or human genes. In the latter case, all phenotype classes associated with any disease that is associated with variants in a gene are assigned to that gene. An example of this is shown in Figure 5.1.

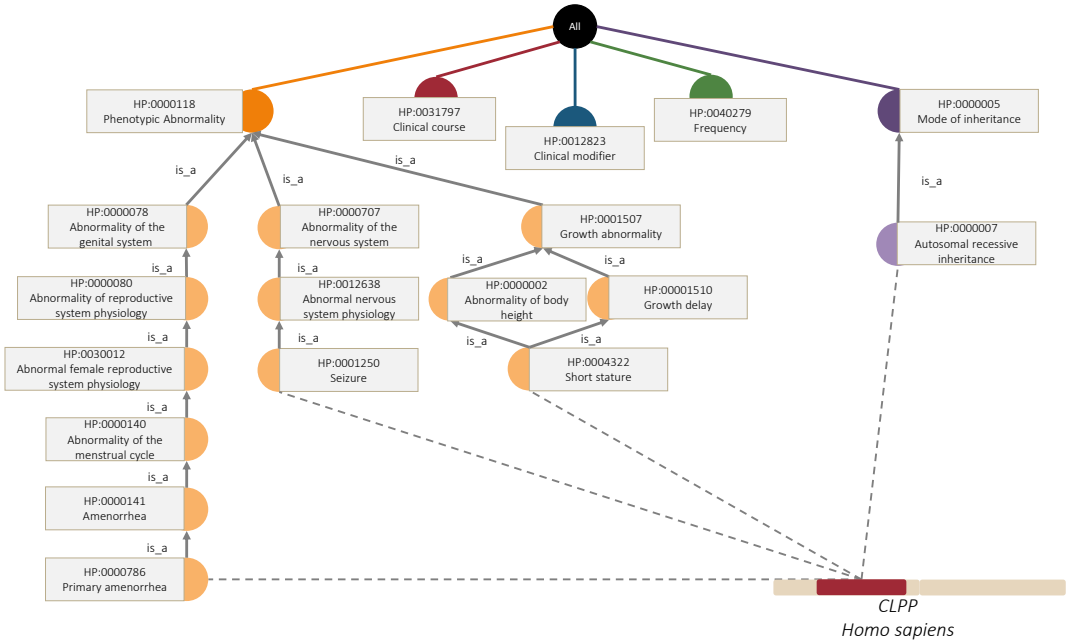


Figure 5.1: Sub-graph of the HPO KG formed by a human gene (*CLPP*) and HPO classes. For this gene, the classes that best describe it are chosen for its annotation.

5.2.2 Proxies of Gene Similarity

For the gene data set, the proxy similarity is based on OMIM's Phenotypic Series (PS) (Amberger et al. [2014]), which group identical or similar phenotypes involved in the same disorder. Up to November 2019, OMIM had 464 different PS comprised of 3,777 phenotypes. Information on how genes relate to PS was retrieved from OMIM[4].

PS similarity (sim_{PS}) is defined as the ratio between the number of PS common to genes A and B and the total number of distinct PS through genes A and B, with

$$sim_{PS}(A, B) = \frac{|PS_A \cap PS_B|}{|PS_A \cup PS_B|} \tag{5.3}$$

[4] https://www.omim.org/phenotypicSeriesTitles/all

where A and B are two genes with the set of PS PS_A and PS_B, respectively.

Similarly to sim_{Pfam} in the MF protein data sets, sim_{Ps} correlates to SS because the more PS two genes are associated with, the more likely it is that they share HPO classes in the "Phenotypic Abnormality" subontology, since PS are a set of similar phenotypes.

5.2.3 Selection of Pairs of Genes for the Benchmark data set

The GP benchmark data set is constituted by genes, identified by their Entrez Gene Code, and annotated with classes from the HPO. The HPO graph (dated November 2019) was collected from the HPO website[5] in OBO format and contains 18,221 ontology classes. HPO annotations were downloaded from the HPO website in a Tab-separated Values (TSV) file (dated November 2019). These annotations link the genes with the HPO classes which best describe the disease in which the genes are proved to play a role. The used KG was made up of 4,293 human genes and its annotations to the HPO. Information on how genes relate to Phenotypic Series (PS) was retrieved from Online Mendelian Inheritance in Man (OMIM)[6] in TSV format.

For the creation of the GP benchmark data set, all human genes meeting the following criteria were considered:

1. The gene must be annotated with at least three classes in the HPO subontology "Phenotypic Abnormality";

2. The gene must have a link with at least one phenotype in any PS, and the mechanism behind that link must be known.

After selecting the eligible entities, pairs of genes were generated to ensure a data set with the same number of pairs of genes with null, not-null and full PS similarity.

5.3 Semantic Similarity Calculation

For all data set pairs, different SSMs were calculated. Each of the selected SSM is a combination of two approaches: the approach used to calculate the IC of an annotating class (IC_{Seco} or IC_{Resnik}) and the IC-based approach used to calculate the similarity between the KG entities (simGIC or BMA). These approaches for IC-based entity similarity were selected because both simGIC and BMA are high-performing classical measures of SS and still widely accepted within the research community, in spite of the new wave of structural-based measures (see Table 3.3).

IC_{Seco}, proposed by Seco et al. [2004], is an intrinsic approach based on the number of direct and indirect children by class c and is given by

$$IC_{Seco}(c) = 1 - \frac{\log[hypo(c) + 1]}{\log[maxnodes]} \tag{5.4}$$

where $hypo(c)$ is the number of direct and indirect children from class c (including class c) and $maxnodes$ is the total number of classes in the ontology.

[5] https://hpo.jax.org/

[6] https://www.omim.org/phenotypicSeriesTitles/all

IC_{Resnik} is a corpus-based/extrinsic approach proposed by Resnik [1995] and based on the number of entities annotated with class c in a KG, with

$$IC_{Resnik}(c) = -\log p(c) \tag{5.5}$$

where $p(c)$ is the probability of annotation in the corpus.

A normalized version of IC_{Resnik} is given by

$$IC_{norm}(c) = \frac{IC_{Resnik}(c)}{\log N} \tag{5.6}$$

with N being the total number of annotations.

Best Match Average (BMA) is a pairwise approach based on the pairwise measure in which the similarity between two classes corresponds to the IC of their most informative common ancestor (Resnik [1995]). In BMA only the best-matching class for each class in each set of classes describing the individuals (i.e. the most similar) is considered to calculate BMA, given by

$$BMA(A,B) = \frac{\sum_{c_1 \in C_A} sim(c_1, c_2)}{2|C(A)|} + \frac{\sum_{c_2 \in C_B} sim(c_2, c_1)}{2|C(B)|} \tag{5.7}$$

where A and B are entities, C is the set of classes c each entity is described with and $sim(c_1, c_2)$ and $sim(c_2, c_1)$ are the highest similarity values found for classes c_1, c_2. The similarity between two classes can be found using Resnik's similarity (Resnik [1995]) given by

$$sim(c_1, c_2) = max(IC(a)) : a \in A(c_1) \cap A(c_2) \tag{5.8}$$

where a is a class in $A(c_i)$, the set of ancestors of c_i.

SimGIC (Pesquita et al. [2007]) is a groupwise approach which resorts to the Jaccard similarity, in which each class c is weighted by its IC. It is given by

$$simGIC(A,B) = \frac{\sum_{c \in C_A \cap C_B} IC(c)}{\sum_{c \in C_A \cup C_B} IC(c)} \tag{5.9}$$

where A and B are entities and C is the set of classes c each entity is annotated with.

By combining the approaches for entity similarity with the different ICs, we arrive at the four state-of-the-art SSMs used for the data sets: BMA_{Resnik}, BMA_{Seco}, $simGIC_{Resnik}$ and $simGIC_{Seco}$. These four measures are representative of the most successful approaches for KG-based SS calculation.

SS between all data set entities was computed using the Semantics Measures Library Toolkit (Harispe et al. [2013]), a Java Toolkit dedicated to semantic measures computation and analysis.

5.4 Automated Evaluation Methods

For the evaluation of SSMs using the benchmark data sets, automated evaluation methods were developed based on the data sets characteristics.

Pearson Correlation Coefficient (PCC) is a measure of the linear correlation between two variables X and Y. The PCC value ranged between -1 and 1, where -1 is total negative linear

correlation, 0 is no linear correlation, and 1 is total positive linear correlation. It is usually represented by r and is given by

$$r_{X,Y} = \frac{\sum_{i=1} n(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1} n(x_i - \bar{x})^2}\sqrt{\sum_{i=1} n(y_i - \bar{y})^2}} \tag{5.10}$$

where n is the sample size, x_i, y_i are the individual sample points indexed with i and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean, analogously for $\bar{y}$.

In the context of evaluating SSMs with similarity proxies, x_i and y_i correspond to the set of SS and proxy similarity values calculated for the n pairs of entities in the benchmark data sets. The PCC for the test SSM should be compared with the PCC for the state-of-the-art SSMs in the same data set, to assess the impact of the test SSM.

The correlation between two similarity measures can also be visualized in a scatter plot, by assessing how well a linear function fits the data. Figure 5.2 shows an overview of r_{XY} values in different scatter plots.

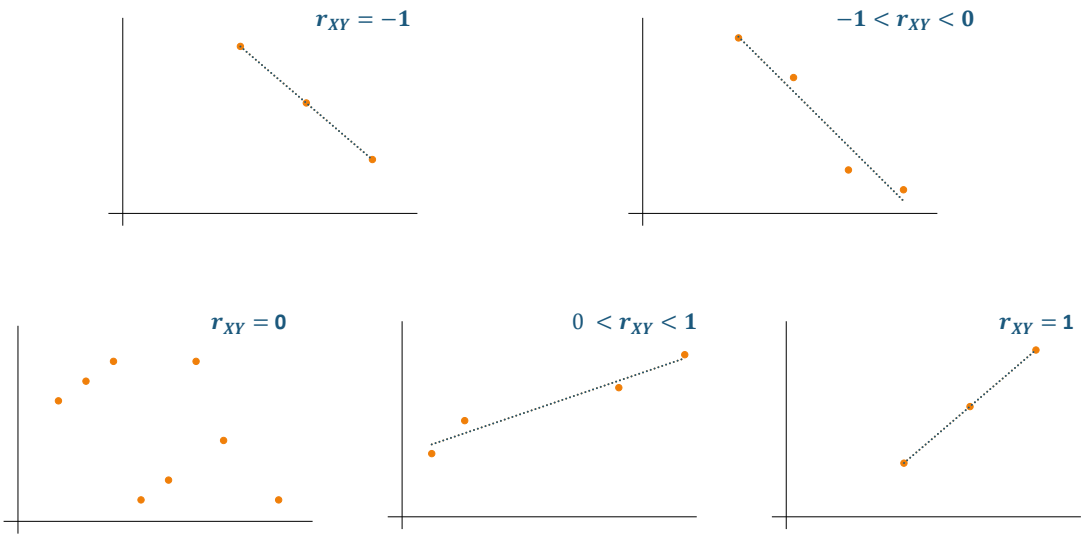


Figure 5.2: Evolution of r_{XY} value, through different scatter plot examples.

For the PPI data sets, SS-based PPI prediction can be employed. This is done by establishing a similarity threshold: if a protein pair has a similarity above the threshold the proteins are considered to interact, otherwise they're not:

$$interaction(A, B) = \begin{cases} \text{interact}, & \text{if } sim(A, B) \geq t \\ \text{don't interact} & \text{otherwise} \end{cases} \tag{5.11}$$

where $sim(A, B)$ is the SS between two proteins A and B and t is the defined threshold for similarity to determine if two proteins interact.

For a given threshold, and knowing the true relation between N protein pairs, it is possible to produce a confusion matrix. A confusion matrix allows for the visualization of the performance of a supervised learning algorithm, in this case SS-based PPI prediction, as depicted in Table 5.2.

Table 5.2: Confusion matrix showing the performance of a supervised learning algorithm, based on the number and type of correct and incorrect predicted labels: true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

		True Interaction label		
		Positive	Negative	Total
Predicted Interaction	Positive	TP	FP	$TP + FP$
	Negative	FN	TN	$FN + TN$
	Total	$TP + FN$	$FP + TN$	N

From the number of TP, FP, TN and FN, we can calculate precision, recall and F_1score as the evaluation scores for the prediction ability of a SSM.

In this case, precision is the fraction of pairs identified as interacting that actually interact and is given by

$$precision = \frac{TP}{TP + FP} \tag{5.12}$$

Recall measures the ratio of interacting pairs correctly identified as such and is given by

$$recall = \frac{TP}{TP + FN} \tag{5.13}$$

F_1score is the measure of the overall accuracy of the algorithm. It is calculated through the harmonic mean of the precision and recall

$$F_1score = 2\frac{precision \cdot recall}{precision + recall} \tag{5.14}$$

Precision, recall and F_1score values vary between 0 and 1, with 1 being the best possible value, and 0 the worst. If the F_1score reaches the value of 1, it means both precision and recall have perfect scores. However, if any of these metrics is 0, F_1score will also be 0 regardless of the other metric score, meaning the algorithm under performs.

Precision-recall plots can be used to visualize evolution of the performance of the algorithm, by varying the similarity threshold values and plotting the resulting precision and recall values.

To determine the best threshold of similarity for a given PPI prediction in a data set, 10-fold cross validation was used. Cross validation is a machine-learning method to evaluate the predictive skill of a classifying algorithm in a data set, as this SS-based PPI prediction algorithm.

In k-fold cross validation, the general procedure is to split the data set into k groups and fit the model k times in $k - 1$ groups, each time leaving one group out as the test set, for the evaluation of the model. In this SS-based PPI prediction algorithm, the procedure is as follows:

1. Shuffle the data set randomly;
2. Split the data set into k groups, k should be an integer respecting the condition $1 < k \leq n$ where n is the data set size;
3. For each group:
   (a) Take the group as the test set;
   (b) Test different similarity threshold values in the remaining $k - 1$ groups;
   (c) Select the threshold that outputs the highest $F_1 score$;
   (d) Evaluate the performance of the selected similarity threshold on the test set.
4. Out of the k similarity thresholds tested in the test sets, select the median similarity threshold.

The best threshold and correspondent precision, recall and $F_1 score$ for all metrics (test and state-of-the-art) can be compared in table or plot format. Alternatively, the performance of all measures can be compared for a selected threshold, not determined by cross validation.

All these evaluation methods were implemented using different Python libraries: Scikit-learn (Pedregosa et al. [2011]) was used to for the evaluation of the SS-based PPI prediction, the plots were designed using Matplotlib (Hunter [2007]), and PCC was calculated resourcing to Pandas (Virtanen et al. [2020]).

Chapter 6

Results and Discussion

This chapter presents the results of the methodology implementation led in Chapter 5, with a discussion of the features, performance and availability of the resulting benchmark.

6.1 Benchmark performance

Applying the proposed methodology to the chosen KGs resulted in 21 different benchmark data sets: ten PPI data sets, ten MF data sets and one GP data set, and associated evaluation techniques. This section goes over the main features of the data sets and the results from linear correlation evaluation techniques, displaying their validity for the proposed application.

The protein data sets comprise pairs of proteins from four different species (*D. melanogaster*, *E. coli*, *H. sapiens*, and *S. cerevisiae*) that support the evaluation of GO-based SSMs, the main application of SS in the biomedical domain, through comparison with relevant biological properties of the proteins (protein sequence, function and interactions). The complete characterization (number of entities, pairs and annotation completeness) of these data sets is presented in tables 6.1 and 6.2.

Table 6.1: Species, number of proteins and pairs and level of annotation completion for all data sets in the PPI collection.

	One Aspect		*All Aspects*	
Species	Proteins	Pairs	Proteins	Pairs
D. melanogaster	481	397	335	270
E. coli	371	738	264	428
H. sapiens	7,644	44,677	7,149	42,204
S. cerevisiae	3,874	34,772	2,959	21,577
All	12,370	80,584	10,707	64,479

Tables 6.3 and 6.4 show PCC between all SSMs and the correspondent data sets similarity proxies. For each data data set, the SSM with the higher PCC for each similarity proxy is highlighted.

Table 6.2: Species, number of proteins and pairs and level of annotation completion for all data sets in the MF collection.

	One Aspect		*All Aspects*	
Species	Proteins	Pairs	Proteins	Pairs
D. melanogaster	7,494	53,795	5,810	52,457
E. coli	1,250	4,623	748	1,813
H. sapiens	13,604	57,906	12,487	57,722
S. cerevisiae	4,783	42,192	3,660	30,747
All	27,131	158,512	22,705	142,736

Table 6.3: PCC between state-of-the-art SSMs (BMA_{Resnik}, BMA_{Seco}, $simGIC_{Resnik}$ and $simGIC_{Seco}$) and sequence similarity (sim_{Seq}) and PPI for all data sets in the PPI collection. SSM with the higher PCC with each similarity proxy is highlighted for each data set.

		One Aspect		*All Aspects*	
Species	Similarity measure	sim_{Seq}	PPI	sim_{Seq}	PPI
D. melanogaster	BMA_{Resnik}	0.401	**0.834**	0.432	**0.795**
	BMA_{Seco}	0.424	0.812	0.464	0.784
	$simGIC_{Resnik}$	0.469	0.691	0.517	0.645
	$simGIC_{Seco}$	**0.483**	0.693	**0.526**	0.647
E. coli	BMA_{Resnik}	0.199	0.700	0.146	0.678
	BMA_{Seco}	0.234	**0.711**	**0.185**	**0.693**
	$simGIC_{Resnik}$	0.219	0.610	0.174	0.595
	$simGIC_{Seco}$	**0.230**	0.626	0.182	0.609
H. sapiens	BMA_{Resnik}	0.400	0.510	0.406	0.510
	BMA_{Seco}	0.385	**0.519**	0.393	**0.520**
	$simGIC_{Resnik}$	**0.545**	0.414	**0.556**	0.413
	$simGIC_{Seco}$	0.535	0.422	0.546	0.421
S. cerevisiae	BMA_{Resnik}	0.240	**0.666**	0.268	**0.642**
	BMA_{Seco}	0.236	0.654	0.269	0.633
	$simGIC_{Resnik}$	**0.300**	0.593	0.349	0.568
	$simGIC_{Seco}$	**0.300**	0.593	**0.351**	0.568
All	BMA_{Resnik}	0.285	0.576	0.312	0.546
	BMA_{Seco}	0.292	**0.583**	0.332	**0.561**
	$simGIC_{Resnik}$	0.372	0.504	**0.435**	0.474
	$simGIC_{Seco}$	**0.374**	0.509	**0.435**	0.479

In Table 6.3, positive correlation in all tests is observed. There is, in most cases, a lower correlation to sequence similarity, as expected (Yu et al. [2010]), with the exception of the *H. sapiens* data sets. BMA is shown to have higher correlation with PPI, while simGIC correlates better to sequence similarity, with both properties being IC independent.

Table 6.4: PCC between state-of-the-art SSMs (BMA_{Resnik}, BMA_{Seco}, $simGIC_{Resnik}$ and $simGIC_{Seco}$) and sequence similarity (sim_{Seq}) and MF similairty (sim_{Pfam}) for all data sets in the MF collection. SSM with the higher PCC with each similarity proxy is highlighted for each data set.

		One Aspect		*All Aspects*	
Species	Similarity measure	sim_{Seq}	sim_{Pfam}	sim_{Seq}	sim_{Pfam}
D. melanogaster	BMA_{Resnik}	0.271	0.324	0.176	0.290
	BMA_{Seco}	0.323	0.404	0.189	0.354
	$simGIC_{Resnik}$	0.493	0.576	0.438	0.609
	$simGIC_{Seco}$	**0.499**	**0.587**	**0.446**	**0.628**
E. coli	BMA_{Resnik}	0.336	0.484	0.347	0.478
	BMA_{Seco}	0.315	**0.485**	0.351	**0.503**
	$simGIC_{Resnik}$	0.340	0.357	0.394	0.367
	$simGIC_{Seco}$	**0.346**	0.378	**0.395**	0.388
H. sapiens	BMA_{Resnik}	0.560	0.586	0.566	0.597
	BMA_{Seco}	0.666	**0.652**	0.680	**0.667**
	$simGIC_{Resnik}$	0.719	0.598	0.729	0.606
	$simGIC_{Seco}$	**0.723**	0.612	**0.732**	0.620
S. cerevisiae	BMA_{Resnik}	0.522	**0.613**	0.512	**0.625**
	BMA_{Seco}	0.588	0.586	0.541	0.605
	$simGIC_{Resnik}$	**0.663**	0.541	**0.662**	0.527
	$simGIC_{Seco}$	0.645	0.555	0.641	0.543
All	BMA_{Resnik}	0.431	0.505	0.388	0.483
	BMA_{Seco}	0.521	0.537	0.490	0.524
	$simGIC_{Resnik}$	**0.573**	0.563	**0.569**	**0.580**
	$simGIC_{Seco}$	0.569	**0.576**	0.566	0.560

In Table 6.4, positive correlations in all tests are also observed. Once more, simGIC correlates better with sequence similarity than any BMA approach and, although this is not true for all data sets, BMA approaches show better correlation with function similarity than simGIC.

The distribution of SSM across all the pairs in the data sets was assessed. For the protein data sets, this was done resorting to each All species *One Aspect* data set, since it contains all the pairs in that proxy's data sets.

Figure 6.1 shows the distribution of all SSMs across all the pairs in the PPI data sets. One should note that a large number of pairs show lower SSM. However, given that from a total of

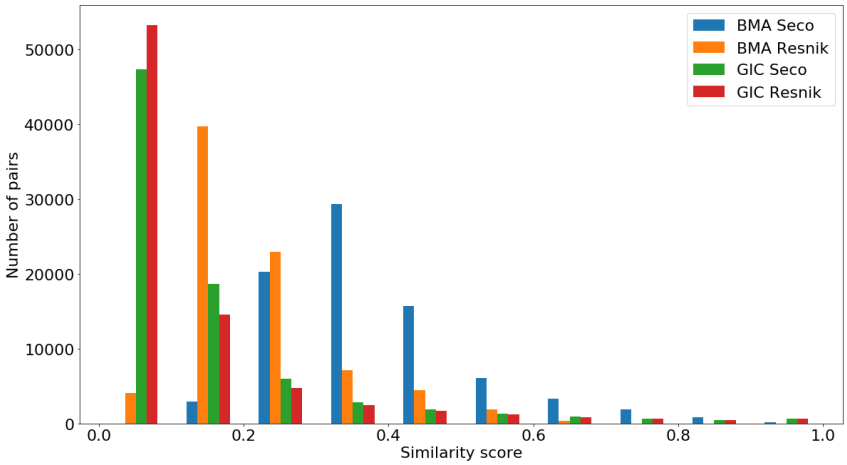


Figure 6.1: Distribution of all SSMs (BMA_{Resnik}, BMA_{Seco}, $simGIC_{Resnik}$ and $simGIC_{Seco}$) values across all species' protein pairs in the PPI data sets.

80,584 pairs of proteins in these data sets, 29,944 have positive interactions, it is expected that the remaining 50,640 to have lower similarity, since they are not reported to interact. These data sets are representative, because there are far more non-interacting proteins in the "real world", however, it is still important to include both positive and negative examples in a benchmark data set.

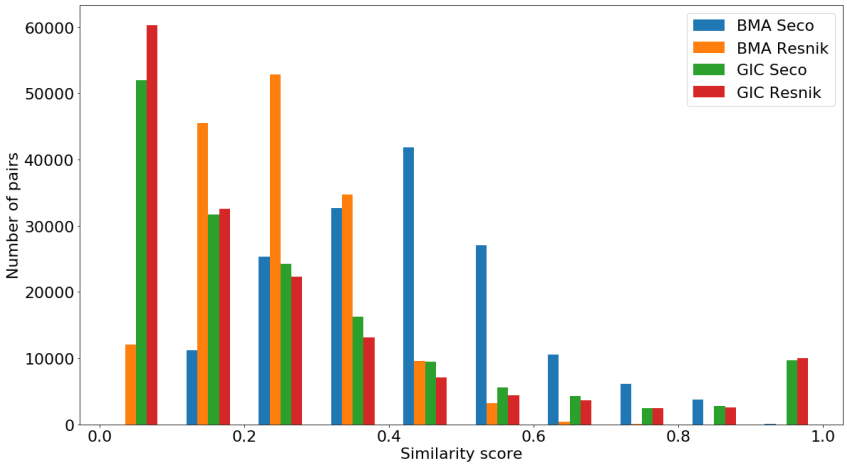


Figure 6.2: Distribution of all SSMs (BMA_{Resnik}, BMA_{Seco}, $simGIC_{Resnik}$ and $simGIC_{Seco}$) values across all species' protein pairs in the MF data sets.

Figures 6.2 and 6.3 show the distribution of similarity across all the pairs in the MF data sets. Analysing these figures, one can see they show a more evenly distributed semantic and MF similarity throughout the pairs. This should be due to having an extra constraint in selecting the pairs for the data sets, based on the MF similarity value, forcing the similarity to be more evenly distributed. Thus, these data sets are representative in terms of values of the different types of similarity measures, following a guideline of major importance for benchmark data sets.

Even though benchmark data sets should be non redundant, the overlap between data sets

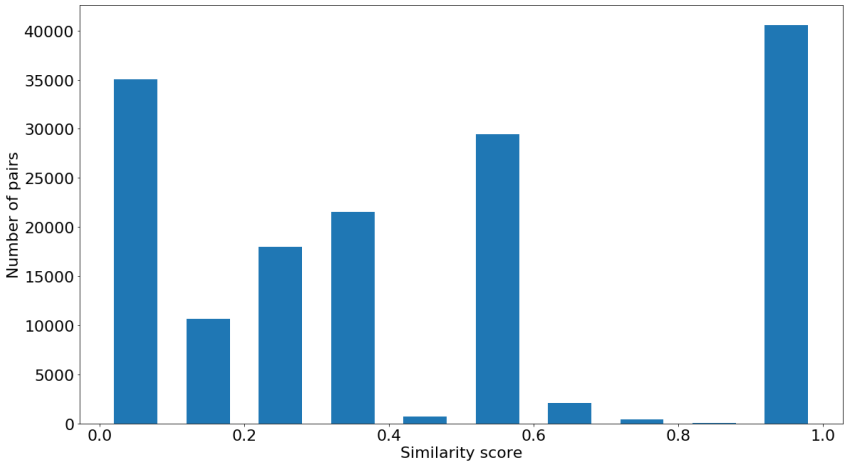


Figure 6.3: Distribution of Sim_{Pfam} values across all species' protein pairs in the MF data sets.

of the same species, but a different level of annotation completion, can be used to evaluate the impact of more thoroughly described proteins in the performance of SSMs. Other levels of annotation completion can be considered in the future, for a different and/or more broad comparison of this impact.

The data sets in each level of annotation completion (*All* data sets) are a compilation of all the protein pairs in each of the species-specific data sets. Even though, once more, there is redundancy between these and the species-specific data sets, the *All* data sets are far larger and can be used for a comparative evaluation of the scalability of the SSMs or SS-based approaches.

Regarding the GP data set, containing pairs of human genes, it supports the evaluation of HPO-based gene SSMs through the role of genes in similar phenotypic disorders. This data set has 12,000 pairs of genes, for a total of 2,026 different gene throughout.

Table 6.5 shows PCC between all SSM and sim_{PS}. The SSM with the higher PCC is highlighted.

Table 6.5: PCC between sim_{PS} and state-of-the-art SSMs (BMA_{Resnik}, BMA_{Seco}, $simGIC_{Resnik}$ and $simGIC_{Seco}$) for the Gene-Phenotypes data set.

	Pearsons correlation coefficient
BMA_{Resnik}	0.572
BMA_{Seco}	**0.590**
$simGIC_{Resnik}$	0.478
$simGIC_{Seco}$	0.482

The SSM that best correlates to sim_{PS} is BMA_{Seco}, however the PCC value for this SSM is similar to those of BMA_{Resnik} and both simGIC approaches, all showing positive correlation between the measures.

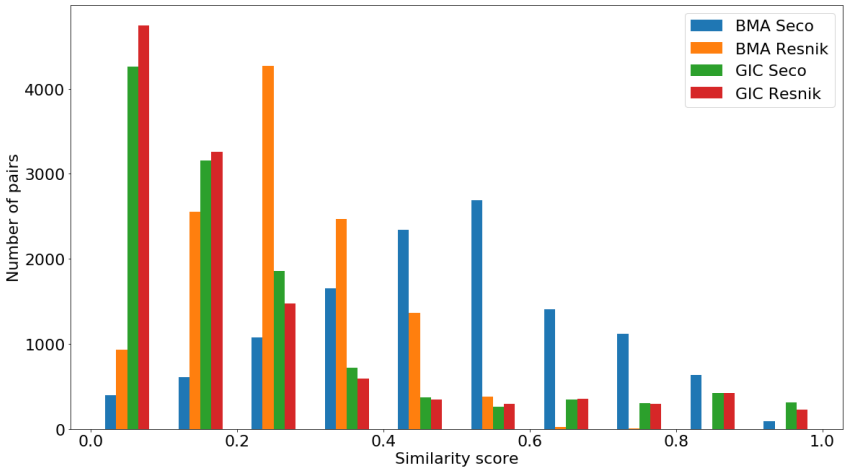


Figure 6.4: Distribution of all SSMs (BMA_{Resnik}, BMA_{Seco}, $simGIC_{Resnik}$ and $simGIC_{Seco}$) values across all the pairs in the GP data sets.

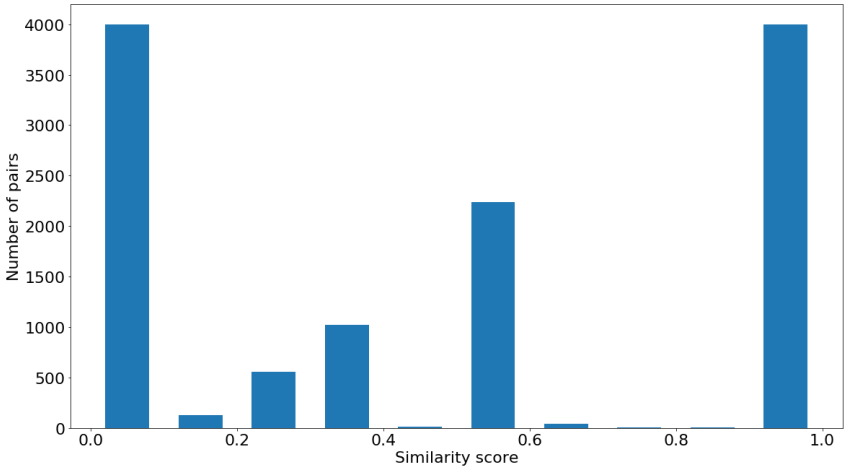


Figure 6.5: Distribution of sim_{PS} values across all the pairs in the GP data sets.

Figures 6.4 and 6.5 show the distribution of similarity across all the pairs in this data set. Similarly to the MF data sets, and due to the same reasons, similarity, both semantic and PS, is fairly distributed throughout the pairs. This means that the data set is representative in terms of values of the different types of similarity measures, which means that it is a good candidate for a SSM evaluation.

6.2 Benchmark availability and usage

This section describes the steps in the usage of the benchmark and it also has some notes on its availability for the research community.

The evaluation of SSMs using the benchmark should follow the steps presented in Figure 4.3. After selecting the measure to evaluate, the SS between the pairs of entities in the chosen data sets must be assessed using that SSM and evaluated by benchmarking its performance against

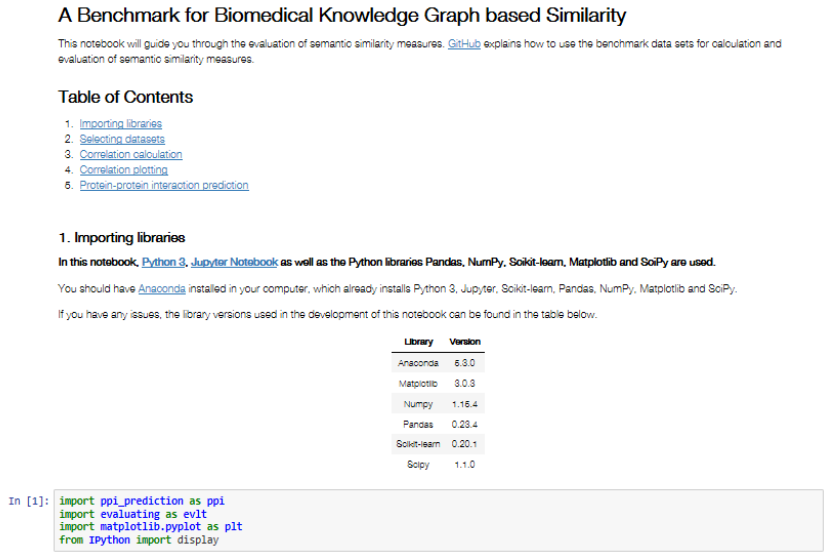
the state-of-the-art SSMs. The benchmark supports different evaluation methods, depending on the similarity proxies available for the used data sets. Table 6.6 shows an overview of the evaluation methods supported for each type of similarity proxy.

Table 6.6: Supported evaluation techniques by each similarity proxy.

Similarity Proxy	Pearson's correlation coefficient calculation	Correlation plotting	Classification performance evaluation
sim_{Seq}	✓	✓	×
sim_{Pfam}	✓	✓	×
PPI	✓	×	✓
sim_{PS}	✓	✓	×

The benchmark sustains a simple evaluation technique, namely the computation and visualization of PCC, but also a more complex one, the PPI prediction evaluation. These are simple evaluation tasks but are relevant for the evaluation of the performance of SSMs and often used when doing so.

The evaluation methods were gathered in an interactive Jupyter Notebook that compiles all the evaluation methods presented in Table 6.6 with a small tutorial on how to perform each evaluation for each data set type. Figure 6.6 shows a screenshot of the introductory section of the Jupyter Notebook and Figure 6.7 shows the results of the evaluation of two arbitrary SSMs (Measure 1 and Measure 2) using a PPI data set.



A Benchmark for Biomedical Knowledge Graph based Similarity

This notebook will guide you through the evaluation of semantic similarity measures. GitHub explains how to use the benchmark data sets for calculation and evaluation of semantic similarity measures.

Table of Contents

1. Importing libraries
2. Selecting datasets
3. Correlation calculation
4. Correlation plotting
5. Protein-protein interaction prediction

1. Importing libraries

In this notebook, Python 3, Jupyter Notebook as well as the Python libraries Pandas, NumPy, Scikit-learn, Matplotlib and SciPy are used.

You should have Anaconda installed in your computer, which already installs Python 3, Jupyter, Scikit-learn, Pandas, NumPy, Matplotlib and SciPy.

If you have any issues, the library versions used in the development of this notebook can be found in the table below.

Library	Version
Anaconda	6.3.0
Matplotlib	3.0.3
Numpy	1.16.4
Pandas	0.23.4
Scikit-learn	0.20.1
Scipy	1.1.0

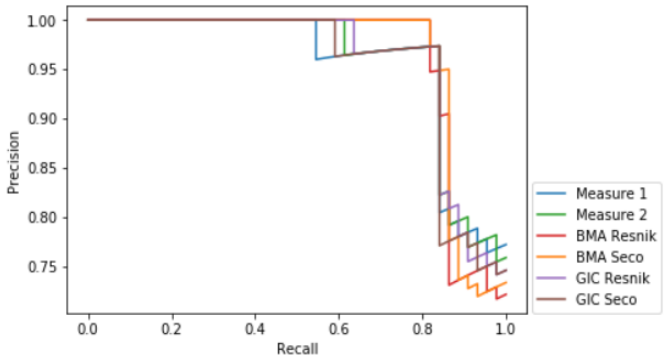
```
In [1]: import ppi_prediction as ppi
        import evaluating as evlt
        import matplotlib.pyplot as plt
        from IPython import display
```

Figure 6.6: Introductory section of the Jupyter Notebook, showing the table of contents and necessary Python libraries for the Notebook use.

	similarity proxy	BMA Resnik	BMA Seco	GIC Resnik	GIC Seco	Measure 1	Measure 2
0	sequence	0.394638	0.475430	0.483673	0.496959	0.501267	0.488958
1	protein protein interaction	0.639708	0.606774	0.470468	0.468584	0.462193	0.465210

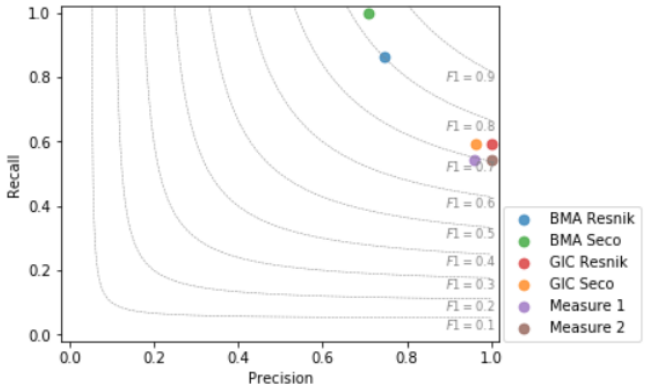
(a) Correlation of the SSMs with similarity proxies.

	metric	Measure 1	Measure 2	BMA Resnik	BMA Seco	GIC Resnik	GIC Seco
0	Best Threshold	0.090000	0.080000	0.280000	0.370000	0.080000	0.090000
1	Precision	0.973684	0.973684	1.000000	0.950000	0.973684	0.973684
2	Recall	0.840909	0.840909	0.818182	0.863636	0.840909	0.840909
3	F1-score	0.902439	0.902439	0.900000	0.904762	0.902439	0.902439



(b) Assessment of the best threshold for each SSM for PPI prediction using cross-validation and Precision-Recall plot.

	metric	BMA Resnik	BMA Seco	GIC Resnik	GIC Seco	Measure 1	Measure 2
0	Precision	0.745098	0.709677	1.000000	0.962963	0.960000	1.000000
1	Recall	0.863636	1.000000	0.590909	0.590909	0.545455	0.545455
2	F-measure	0.800000	0.830189	0.742857	0.732394	0.695652	0.705882



(c) Assessment of PPI prediction ability of the SSMs at a given threshold for SS.

Figure 6.7: Example of the results produced by the Jupyter Notebook when evaluating two arbitrary measures (Measure 1 and Measure 2) using a PPI data set.

The 21 data sets and the developed Jupyter Notebook were made publicly available to the research community[1]. In addition, all the data that was used to compute the SS for all data sets is made available in two different formats (1) Separate files for the ontologies and the annotations files used (2) KGs, in OWL format, containing both the ontology and the instances (divided by species). This allows for a direct comparison with the pre-computed SSMs in the data sets, as well as facilitates the direct comparison between different SSM evaluation works that use this resource, without needing to implement and/or compute the results. For this reason, the benchmark will purposefully remain static for a few years, following the approach used by CESSM (Pesquita et al. [2009b]), released in 2009 and updated in 2014.

The availability and accessibility of this resource promotes good practice in data management and publication, by accounting for reusability. This ensures that the main goal of this benchmark is achieved, by providing data and tools for the interoperable research of SS-based applications and evaluation, and make these studies directly comparable. This means that the benchmark is in line with the FAIR guiding principles.

6.3 Discussion

A big issue in the evaluation of SSMs is the diversity in the studies employed to do so. SSMs are usually tested in a small and controlled set of data, developed for that study alone. This unsystematic assessment practice can lead to biases in the published results, especially if not compared with those of the state-of-the-art SSMs in the same conditions, i.e. using the exact same version of the KG and the same entity pairs. Moreover, not employing a common strategy or, at least, the same data, makes the results from these studies not directly comparable across them.

This benchmark aimed at tackling these issues by providing data sets with pairs of entities of different species, annotated with different ontologies and providing a combination of different similarity proxies and multiple state-of-the-art SSMs. In fact, the benchmark represents an evolution compared with the previous efforts in this area, both in terms of the size and diversity of the data employed. This can be seen in Table 6.7, when comparing these data sets to both updates of CESSM (Pesquita et al. [2009b]).

Table 6.7: Comparison between CESSM and KG Sim Benchmark in terms of number of entities, pairs, ontologies and species.

Resource	CESSM 2009	CESSM 2014	KG Sim Benchmark
Entities	1,039	1,626	30,831
Pairs	13,430	22,302	249,875
Ontologies	GO	GO	GO & HPO
Species	Non-specific		4

Using multiple KGs benefits benchmark utility due to the different structure and context of

[1] https://github.com/liseda-lab/kgsim-benchmark

these graphs. Although both are biomedical ontologies, the GO provides a structured vocabulary for the annotation of gene products regarding their role in the cell and organism, and the HPO is an ontology of medically relevant phenotypes. The difference in the structure of these two ontologies can also play a role in the performance of the measure. The GO is divided in three independent subontologies that capture different aspects of a gene product's role in the cell, all of them necessary for a full characterization of the gene product. The HPO, on the other hand has five subontologies, but only the main one ("Phenotypic Abnormality") was considered when calculating the state-of-the art similarity for the HPO data set. Evaluating a SSM with these two different KG could yield different results for the performance of the measure, due to the highlighted differences between them. This suggests that testing the same SSM in differently targeted KGs can be a good evaluator of its ability to generalize to different KGs, entity types and applications.

In order to guarantee that the SSMs can capture the functional similarity between the entities, their meaning must be well captured within the ontology context. This meant selecting entities annotated with more specific ontology classes (classes with fewer child classes), as sharing one, or more, of these classes will result in a higher and more significant SS between the two entities. This was done in order to tackle the shallow annotation problem for SSMs, that results in SS values that are inconsistent with human perception, due to shallowly described entities (Pesquita [2017]).

However, there is still another feature that can impact SS results, the annotation length. The annotations each entities carries can impact SS values. Annotations are not uniformly distributed among entities: some entities are widely studied and are very well annotated, while others are poorly described. This can hinder the performance of SSMs in tasks such as PPI prediction (Jiang et al. [2014]). There is, additionally, a positive correlation between SS and the number of annotations from a pair of entities, whereas as the difference in the number of annotations between the entities increases, their SS can decrease (Kulmanov and Hoehndorf [2017]).

Thus, it was non-trivial to establish the criteria for entity selection, to avoid both these issues. The selected criteria established a minimum number of annotation per entity, but no maximum number or threshold for annotation size difference. However, the selected entities are well characterized with other biological data (e.g. sequence, functional domains, involvement in diseases), which suggests that the entities and pairs in the data sets should have a balanced number of annotations. Nonetheless, the data sets can be employed for studying the existence of this effect in a new set of data sets, with the selected SSMs. Furthermore, the developed methodology can be easily applied to generate benchmarks targeting different criteria for entity pair selection, such as those based on annotation length, to produce tailored data sets to investigate this specific issue.

The benchmark takes advantage of proxies of entity similarity for the evaluation of SSMs as a device for determining functional similarity of two biomedical entities. The definition of biological function similarity is ambiguous because its exact meaning varies based on the context in which it is used (Friedberg [2006]). This bias is especially relevant when similarity is being defined by domain experts. For instance, let us imagine two protein kinases. These are proteins

that modify other proteins by adding a phosphate group to them. A biochemist could deem the two proteins as very similar because they're both kinases, therefore, they have the same function. However, when analysing the two proteins from a physician's perspective, they might be more interested in the role these two proteins play in the whole-organism level. The two kinases may be involved in different signaling pathways, and different mutations in this kinases might cause different diseases. Thus, from a physiological point of view, the two kinases are dissimilar. Not only is it unfeasible to ask domain experts to do this manual verification of similarity for every pair of biological entities there is, due to the amount of data in these domains, their perception will always be biased to their field of study or area of expertise.

Enter similarity proxies. These measures of similarity, despite still capturing only one functional aspect of the entities at a time, bear two advantages: they rely on objective representations of the entities (e.g. gene sequence, protein structure, existence of PPI, metabolic pathways affected by the disease) and calculate similarity using mathematical expressions or other algorithms. Not only can these algorithms compare entities at a much faster rate than human experts, they can quantify the result from that comparison, as opposed to a similar/dissimilar assessment.

The benchmark supports the evaluation of SSMs based on four different similarity proxies: protein sequence similarity, existence of PPIs, MF similarity and PS similarity. The first three proxies are proxies for protein similarity and can be used to evaluate GO-based SSMs, whereas the latter is a proxy for gene similarity, for the evaluation of HPO-based SSMs.

There was a bigger effort towards developing the evaluation methods for GO-based SSMs, because GO is the most widely used ontology in the study of SSM and its applications. The GO-based benchmark data sets can be divided by the similarity proxies employed in them. For each of the data sets the combination of similarity proxies can either be (1) PPIs and sequence similarity or (2) MF and sequence similarity. Sequence similarity is considered for both these data sets because, not only can it be computed for any two proteins for which sequence is known, and there is no restriction in the pairs of proteins for which it is computed, but also because sequence similarity does not show a strong enough relation with SS (Ikram et al. [2018]) to be used alone, as a sole evaluator of SSMs. As exposed in Chapter 5, PPI and MF are known to have different relation with the GO aspects. While MF similarity is expected to correlate better with more matching classes from the MF subontology, the existence of a PPI is more likely to be in agreement with overlapping classes in the CC and BP subontologies. Thus, a SSM that has a positive relation with both these proxies, is a SSM that does a good job in capturing entity similarity, as it is capable of considering different aspects of it.

The HPO-based data set considers only one similarity proxy, PS similarity. This similarity proxy can be seen as an evaluation of how well SS captures the probability of two genes being involved in the same disorders.

Finally, the selected state-of-the-art SSMs, BMA_{Resnik}, BMA_{Seco}, $simGIC_{Resnik}$ and $simGIC_{Seco}$, are classic approaches for SS made up of a combination of an approach to calculate the IC of the ontology classes that describe the entities, and an approach for combining those values of IC to achieve a value for the SS between the two entities. By combining the two different-natured approaches of each component of a SSM we capture a full spectrum of successful

approaches for the calculation of SS. This provides a broader term of comparison for the SSMs tested using the benchmark, as opposed to providing the results using only one type of state-of-the-art SSMs.

Through the analysis of Tables 6.3 and 6.4 simultaneously, it is possible to see that simGIC is the measure with the best correlation with sequence similarity, in spite of the species, level of annotation and the other similarity proxy considered (MF similarity or PPI). This is consistent with other observations, where simGIC ranks in the top in terms of correlation with sequence similarity (Wu et al. [2013]; Pesquita et al. [2008]). The correlation between SS and MF similarity is overall better when using BMA, although being sometimes outperformed by simGIC. Most research indicates a similar performance by both these methods in different data sets (Bible et al. [2017]; Dutta et al. [2017]; Wu et al. [2013]). In terms of relation to PPIs, BMA outperforms simGIC in all data sets. Once more, this was expected, following previous evaluation comparing both measures (Jain and Bader [2010]). The Maximum approach is usually better than BMA in this task because interacting proteins only need to share a CC or BP class for their similarity to be biologically relevant for PPIs. Nonetheless, BMA still performs well and is used in more SS-based applications, while the Maximum approach is unsuitable to assess global similarity (Pesquita [2017]), thus the decision of using BMA instead of the Maximum approach. Finally, Table 6.5 shows that PS similarity correlates well with all four SSMs used, its behaviour being similar to that of MF similarity, probably due to being calculated using the same formula.

In sum, the state-of-the-art SSMs perform well in all correlation tests they were subjected to, meaning these results are in line with other assessments made in the past.

The features of the benchmark data sets in here discussed, along with making the benchmark easy to obtain and use, show that the benchmark follows the guidelines that should be considered in building a trust-worthy benchmark.

Chapter 7

Conclusion

The diversity in applications for SS, especially when studying biological entities, has widely boosted the development of new SSMs that formalize the notion of similarity in slightly different ways and may disagree on what makes two entities similar or distinct. This variety in measures for calculating SS, coupled with the diversity in applications that benefit from it, raises the need to answer the question: *"what is the best SSM for each application?"*.

Since there is no gold standard for similarity between biological entities, because these are very complex entities, one solution is to compare the SSMs to measures, or proxies, of biological similarity, and assess how the SSMs capture the entities similarity from different lenses.

The goal of this dissertation was to develop a benchmark for SSMs in the biomedical domain, that supports the large-scale evaluation of SSMs, by exploiting proxies for biomedical entity similarity. The benchmark is made up of a collection of benchmark data sets and automated evaluation techniques for SSMs. This was done by developing and applying a methodology for the construction of benchmark data sets for KG based SS and fitting evaluation techniques, supported by the features of the data sets and their underlying entities.

The methodology developed within this dissertation was applied to two KGs: proteins annotated with the GO and genes annotated with the HPO, and resulted in a benchmark made up of a collection of 21 benchmark data sets and evaluation techniques that fit the data sets characteristics.

Out of the 21 data sets, 20 are based on the GO KG, given it is the most successful case of the use of an ontology in the biomedical domain and it supports the annotation of gene products of different species. There are two major groups of data sets: one based on MF similarity and another based on PPIs. These proxies of similarity capture the similarity of entities through different lenses and can evaluate how well the different SSMs fit them. These data sets are also divided by species and level of annotation completion. The diversity in these features among the data sets can also have an impact in the performance of a SSM.

The HPO data set exploits the role of human genes in different phenotypic abnormalities. This data set, along with the evaluation of SSMs, can also be used to evaluate the impact of SS in the prediction of genes that play a role in the same diseases, since if PS similarity is not null, the genes will be involved in, at least, one matching disorder.

The collection of data sets also represents a contribution in itself, since the data sets therein

can be used for the employment of other SS-based approaches, as supervised/unsupervised learning techniques. It also represents an evolution compared with the previous efforts in this area, both in terms of the size and diversity of the data employed.

The benchmark supports the calculation of correlation between the different SSMs and the similarity proxies of each data set. Additionally, for the PPI data sets, it also supports the evaluation of the predictive power of SSMs when performing SS-based PPI prediction. The benchmark data can also be used to evaluate the multiple components of a SSM, i.e., IC, class-based similarity, and instance-based similarity approaches.

Overall, the benchmark follows the FAIR guiding principles for scientific data management and stewardship. The data sets use the adequate URIs for all entities in the data sets and KG data used, the benchmark is available and accessible for the use by the research community and, by providing the relevant KG data used to produce the data set, its reusability is promoted and encouraged. Ultimately, the goal of this benchmark is to lead to interoperability of the results in the evaluation of SSM, for direct comparisons between different studies.

Despite being domain-specific, it is expected that this benchmark and collection of data sets to also be useful beyond the biomedical domain. Similarity computation within KGs is a fundamental building block of many semantic web applications ranging from data integration to data mining, meaning the benchmark data sets can be used for the evaluation of SSMs developed outside the biomedical domain.

7.1 Future Work

This dissertation presented a benchmark for the large scale evaluation of SSMs in the biomedical domain. However, it can still be updated to improve the existing features or add new ones. In the future, updates to the benchmark can include:

- Updates to the existing data sets: inclusion of other state-of-the-art SSMs, inclusion of new similarity proxies and evaluation techniques, updated selection features for the entities, etc;
- Inclusion of data sets with pairs of entities annotated with different biomedical ontologies;
- Expansion of the GO data set collection: inclusion of pairs of proteins of other species, inclusion of SS calculated under each aspect of GO to support other learning techniques, etc.
- Expansion of the HPO data set collection: inclusion of data sets with pairs of human diseases, to support the evaluation of SSM in the prediction of diagnosis for patients.

Furthermore, the general approach developed for the creation of the data sets is generalizable to any domain where a similarity proxy can be created, making the development of analogous benchmarks outside the biomedical domain a possibility.

References

Al-Mubaid, H. and Nagar, A. (2008). Comparison of four similarity measures based on GO annotations for Gene Clustering. *2008 IEEE Symposium on Computers and Communications*, pages 531–536. 15, 17

Ali, W. and Deane, C. M. (2009). Functionally guided alignment of protein interaction networks for module detection. In *Bioinform.* 17

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2014). Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1):D789–D798. 29

Aniba, M., Poch, O., and Thompson, J. (2010). Issues in bioinformatics benchmarking: The case study of multiple sequence alignment. *Nucleic acids research*, 38:7353–63. 11

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A. P., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25:25–29. 2, 25

Bairoch, A. (2000). The enzyme database in 2000. *Nucleic acids research*, 28(1):304–305. 18

Bandyopadhyay, S. and Mallick, K. (2013). A new path based hybrid measure for gene ontology similarity. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 11. 16

Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706–716. 1

Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl_1):i38–i46. 28

Benabderrahmane, S., Smaïl-Tabbone, M., Poch, O., Napoli, A., and Devignes, M.-D. (2010). Intelligo: a new vector-based semantic similarity measure including annotation origin. In *BMC Bioinformatics*. 17

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic Web. *Scientific American*, 284(5):28–37. 5

Bible, P., Sun, H.-W., Morasso, M., Loganantharaj, R., and Wei, L. (2017). The effects of shared information on semantic calculations in the gene ontology. *Computational and Structural Biotechnology Journal*, 15. 46

Bodenreider, O., Aubry, M., and Burgun-Parenthoine, A. (2004). Non-lexical approaches to identifying associative relations in the gene ontology. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 91–102. 16

Bodenreider, O. and Stevens, R. (2006). Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, 7(3):256–274. 5

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795. 19

Chabalier, J., Mosser, J., and Burgun, A. (2007). A transversal approach to predict gene product networks from ontology-based similarity. *BMC bioinformatics*, 8:235. 17

Cheatham, M. and Hitzler, P. (2014). Conference v2. 0: An uncertain version of the oaei conference benchmark. In *International Semantic Web Conference*, pages 33–48. Springer. 17

Chen, X., Yang, R., Xu, J., Hongzhe, M., Chen, S., Bian, X., and Liu, L. (2012). A sensitive method for computing go-based functional similarities among genes with 'shallow annotation'. *Gene*, 509:131–5. 17

Cheol Jeong, J. and Chen, X. (2015). A new semantic functional similarity over gene ontology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(2):322–334. 16

Cho, Y.-R., Hwang, W., Ramanathan, M., and Zhang, A. (2007). Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC Bioinformatics*, 8:265 – 265. 17

Consortium, T. U. (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515. 11, 27

Couto, F. M., Silva, M. J., and Coutinho, P. (2007). Measuring semantic similarity between gene ontology terms. *Data Knowl. Eng.*, 61:137–152. 16

Couto, F. M., Silva, M. J., and Coutinho, P. M. (2003). Implementation of a functional semantic similarity measure between gene-products. 16

Dutta, P., Basu, S., and Kundu, M. (2017). Assessment of semantic similarity between proteins using information content and topological properties of the gene ontology graph. *IEEE/ACM transactions on computational biology and bioinformatics*, PP. 46

Dörpinghaus, J. and Jacobs, M. (2019). Semantic knowledge graph embeddings for biomedical research: Data integration using linked open data. 8

Ehsani, R. and Drablos, F. (2016). Topoicsim: A new semantic similarity measure based on gene ontology. *BMC Bioinformatics*, 17. 16

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432. 18, 26

Ferreira, J. and Couto, F. (2010). Semantic similarity for automatic classification of chemical compounds. *PLoS computational biology*, 6. 16

Ferreira, J. and Couto, F. (2019). Multi-domain semantic similarity in biomedical research. *BMC Bioinformatics*, 20:246. 17

Friedberg, I. (2006). Automated protein function prediction—the genomic challenge. *Briefings in Bioinformatics*, 7(3):225–242. 44

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S. M., Irizarry, R. A., Leisch, F., Li, C., Mächler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G. K., Tierney, L., Yang, J. Y., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80 – R80. 17

Golbreich, C., Horridge, M., Horrocks, I., Motik, B., and Shearer, R. (2007). Obo and owl: Leveraging semantic web technologies for the life sciences. In Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., and Cudré-Mauroux, P., editors, *The Semantic Web*, pages 169–182, Berlin, Heidelberg. Springer Berlin Heidelberg. 7

Gong, X., Jiang, J., Duan, Z., and Lu, H. (2018). A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology. *BMC bioinformatics*, 19(Suppl 4):162. 15

Guzzi, P. H., Mina, M., Guerra, C., and Cannataro, M. (2011). Semantic similarity analysis of protein data: assessment with biological features and issues. *Briefings in Bioinformatics*, 13(5):569–585. 16

Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2013). The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, 30(5):740–742. 31

Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2015). Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, 8(1):1–254. 2

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., and Steinbeck, C. (2016). Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44(D1):D1214—9. 7, 15

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2011). PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 39(18):e119–e119. 15

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2015). The role of ontologies in biological and biomedical research: a functional perspective. In *Briefings in Bioinformatics*. 2, 5, 10

Horrocks, I. (2008). Ontologies and the semantic web. *Communications of the ACM*, 51:58–67. 5, 6, 7

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J. A., Stephens, R. M., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2007). The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8:R183 – R183. 17

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95. 34

Hunter, L. (2017). Knowledge-based biomedical data science. *Data Science*, 1:1–7. 8

Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J., and O'donovan, C. (2014). The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Research*, 43(D1):D1057–D1063. 25

Ikram, N., Qadir, M. A., and Afzal, M. T. (2018). Investigating Correlation between Protein Sequence Similarity and Semantic Similarity Using Gene Ontology Annotations. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 15(3):905–912. 26, 45

Jacob, E. (2005). Ontologies and the semantic web. *Bulletin of the American Society for Information Science and Technology*, 29:19 – 22. 5

Jain, S. and Bader, G. D. (2010). An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, 11(1):562. 15, 16, 28, 46

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *ROCLING*. 16

Jiang, Y., Clark, W. T., Friedberg, I., and Radivojac, P. (2014). The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective. *Bioinformatics*, 30(17):i609–i616. 44

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., and et al. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464. 7, 15

Kulmanov, M. and Hoehndorf, R. (2017). Evaluating the effect of annotation size on measures of semantic similarity. *Journal of Biomedical Semantics*, 8. 44

Kustra, R. and Zagdanski, A. (2006). Incorporating Gene Ontology in Clustering Gene Expression Data. In *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pages 555–563. 15

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., and et al. (2018). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, 47(D1):D1018–D1027. 15, 28, 29

Lee, H. K., Hsu, A. K.-H., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome research*, 14 6:1085–94. 17

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195. 1

Li, B., Luo, F., Wang, J. Z., Feltus, F. A., and Zhou, J. (2010). Effectively integrating information content and structural relationship to improve the go-based similarity measure between proteins. In *BIOCOMP*. 16

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 16

Liu, W., Liu, J., and Rajapakse, J. C. (2018). Gene ontology enrichment improves performances of functional similarity of genes. *Scientific Reports*, 8(1):12100. 19

Liu, Z.-P., Wu, L.-Y., Wang, Y., Chen, L., and Zhang, X. (2007). Predicting gene ontology functions from protein's regional surface structures. *BMC bioinformatics*, 8:475. 15

Maetschke, S. R., Simonsen, M., Davis, M. J., and Ragan, M. A. (2011). Gene ontology-driven inference of protein–protein interactions using inducers. *Bioinformatics*, 28(1):69–75. 26, 28

Mahdavi, M. and Lin, Y.-H. (2007). False positive reduction in protein-protein interaction predictions using gene ontology annotations. *BMC bioinformatics*, 8:262. 15

Makrodimitris, S., Ham, R., and Reinders, M. (2018). Improving protein function prediction using protein sequence and go-term similarities. *Bioinformatics (Oxford, England)*, 35. 15

Mangul, S., Martin, L., Hill, B., Lam, A., Distler, M., Zelikovsky, A., Eskin, E., and Flint, J. (2019). Systematic benchmarking of omics computational tools. *Nature Communications*, 10(1). 2, 11

Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., and Jacq, B. (2004). Gotoolbox: functional analysis of gene datasets based on gene ontology. *Genome Biology*, 5:R101 – R101. 17

Masino, A., Dechene, E., Dulik, M., Wilkens, A., Spinner, N., Krantz, I., Pennington, J., Robinson, P., and White, P. (2014). Clinical phenotype-based gene prioritization: An initial

study using semantic similarity and the human phenotype ontology. *BMC bioinformatics*, 15:248. 15

Mesquita da Costa, M., Reeve, S., Grumbling, G., and Osumi-Sutherland, D. (2013). The drosophila anatomy ontology. *Journal of biomedical semantics*, 4:32. 7

Mistry, M. and Pavlidis, P. (2008). Gene ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 9:327 – 327. 17

Morales, C., Collarana, D., Vidal, M.-E., and Auer, S. (2017). Matetee: A semantic similarity metric based on translation embeddings for knowledge graphs. In *International Conference on Web Engineering*, pages 246–263. Springer. 18

Mortensen, J. M., Musen, M. A., and Noy, N. F. (2013). Crowdsourcing the verification of relationships in biomedical ontologies. In *AMIA Annual symposium proceedings*, volume 2013, page 1020. 17

Othman, R. M., Deris, S., and Illias, R. M. (2008). A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. *Journal of biomedical informatics*, 41 1:65–81. 16

Palma, G., Vidal, M.-E., Haag, E., Raschid, L., and Thor, A. (2015). Determining similarity of scientific entities in annotation datasets. *Database*, 2015. 18

Paul, M. and Anand, A. (2018). A New Family of Similarity Measures for Scoring Confidence of Protein Interactions using Gene Ontology. 18

Pedersen, T., Pakhomov, S. V., Patwardhan, S., and Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299. 18

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. 34

Pekar, V. and Staab, S. (2002). Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In *COLING*. 16

Pesaranghader, A., Matwin, S., Sokolova, M., and Beiko, R. G. (2015). simDEF: definition-based semantic similarity measure of gene ontology terms for functional similarity analysis of genes. *Bioinformatics*, 32(9):1380–1387. 16

Pesquita, C. (2017). Semantic similarity in the gene ontology. In Dessimoz, C. and Škunca, N., editors, *The Gene Ontology Handbook*, pages 161–173. Humana Press, New York, NY, USA. 16, 22, 26, 44, 46

Pesquita, C., Faria, D., Bastos, H., Falcão, A., and Couto, F. (2007). Evaluating go-based semantic similarity measures. *Proc 10th Annual Bio-Ontologies Meeting*. 31

Pesquita, C., Faria, D., Bastos, H., Ferreira, A., Falcão, A., and Couto, F. (2008). Metrics for go based protein semantic similarity: A systematic evaluation. *BMC bioinformatics*, 9 Suppl 5:S4. 17, 26, 46

Pesquita, C., Faria, D., Falcão, A. O., Lord, P. W., and Couto, F. M. (2009a). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7). 8, 9

Pesquita, C., Pessoa, D., Faria, D., and Couto, F. (2009b). Cessm: Collaborative evaluation of semantic similarity measures. *JB2009: Challenges in Bioinformatics*, 157. 18, 43

Popescu, M., Keller, J. M., and Mitchell, J. A. (2006). Fuzzy measures on the gene ontology for gene product similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3:263–274. 17

Pulido, J., Ruiz, M., Herrera, R., Cabello, E., Legrand, S., and Elliman, D. (2006). Ontology languages for the semantic web: A never completely updated review. *Knowledge-Based Systems*, 19(7):489 – 497. Creative Systems. 5

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 16, 31

Ristoski, P., de Vries, G. K. D., and Paulheim, H. (2016). A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In *The Semantic Web – ISWC 2016*, pages 186–194, Cham, Switzerland. Springer International Publishing. 19

Rosse, C. and Mejino, J. (2004). A reference ontology for biomedical informatics: The foundational model of anatomy. *Journal of biomedical informatics*, 36:478–500. 7

Sanfilippo, A., Posse, C., Gopalan, B., Riensche, R., Beagley, N., Baddeley, B., Tratz, S., and Gregory, M. (2007). Combining hierarchical and associative gene ontology relations with textual evidence in estimating gene and gene product similarity. *IEEE Transactions on NanoBioscience*, 6:51–59. 16

Sarkar, A., Yang, Y., and Vihinen, M. (2020). Variation benchmark datasets: update, criteria, quality and applications. *Database*, 2020. baz117. 12

Schlicker, A., Domingues, F. S., Rahnenführer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7:302 – 302. 16

Seco, N., Veale, T., and Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI'04, pages 1089–1090, Amsterdam, The Netherlands, The Netherlands. IOS Press. 30

Sheehan, B., Quigley, A. J., Gaudin, B., and Dobson, S. A. (2008). A relation based measure of semantic similarity for gene ontology annotations. *BMC Bioinformatics*, 9:468 – 468. 17

Socher, R., Chen, D., Manning, C. D., and Ng, A. (2013). Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934. 19

Sousa, R. T., Silva, S., and Pesquita, C. (2020). Evolving knowledge graph similarity for supervised learning in complex biomedical domains. *BMC bioinformatics*. 26

Stevens, R., Wroe, C., Lord, P., and Goble, C. (2004). Ontologies in bioinformatics. In *Handbook on Ontologies*. Springer, Berlin, Heidelberg. 1, 7

T. P., A. (2011). Bioinformatics-the explosion of modern science and technology. *Drug Invention Today*, 3:262–264. 1

Tan, F., Yang, R., Xu, X., Chen, X., Wang, Y., Hongzhe, M., Liu, X., Wu, X., Chen, Y., Liu, L., and Jia, X. (2014). Drug repositioning by applying 'expression profiles' generated by integrating chemical structure similarity and gene semantic similarity. *Molecular bioSystems*, 10. 16

Teng, Z., Guo, M., Liu, X., Dai, Q., Wang, C., and Xuan, P. (2013). Measuring gene functional similarity based on group-wise comparison of go terms. *Bioinformatics (Oxford, England)*, 29. 17

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. . . (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272. 34

Wang, H., Azuaje, F., and Bodenreider, O. (2005). An ontology-driven clustering method for supporting gene expression analysis. In *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, pages 389–394. 15

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23 10:1274–81. 16

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. (2011a). Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(suppl_2):W541–W545. 1

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. (2011b). BioPortal: enhanced functionality via new web services from the

national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(suppl_2):W541–W545. 7

Wilkinson, M., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L. O., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., and Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3. 10

Wu, H., Su, Z., Mao, F., Olman, V., and Xu, Y. (2005). Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic Acids Research*, 33:2822 – 2837. 16, 17

Wu, X., Pang, E., Lin, K., and Pei, Z.-M. (2013). Improving the measurement of semantic similarity between gene ontology terms and gene products: Insights from an edge- and ic-based hybrid method. *PloS one*, 8:e66745. 16, 46

Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *ACL*. 16

Xu, Y., Guo, M.-z., Shi, W., Liu, X., and Wang, C. (2013). A novel insight into gene ontology semantic similarity. *Genomics*. 16

Xue, H., Peng, J., and Shang, X. (2019). Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on hpo. *BMC Systems Biology*, 13. 15

Ye, P., Peyser, B. D., Pan, X., Boeke, J. D., Spencer, F. A., and Bader, J. S. (2005). Gene function prediction from congruent synthetic lethal interactions in yeast. *Molecular Systems Biology*, 1:2005.0026 – 2005.0026. 17

Yu, H., Gao, L., Tu, K., and Guo, Z. (2005). Broadly predicting specific gene functions with expression similarity and taxonomy similarity. *Gene*, 352:75–81. 16

Yu, H., Jansen, R., and Gerstein, M. (2007). Developing a similarity measure in biological function space. 17

Yu, J., Guo, M., Needham, C. J., Huang, Y., Cai, L., and Westhead, D. R. (2010). Simple sequence-based kernels do not predict protein–protein interactions. *Bioinformatics*, 26(20):2610–2614. 28, 37

Yu, Z., Luo, W., Fu, G., and Wang, J. (2016). Interspecies gene function prediction using semantic similarity. *BMC Systems Biology*, 10:495–507. 15

Zhang, J., Jia, K., Jia, J., and Qian, Y. (2018). An improved approach to infer protein-protein interaction based on a hierarchical vector space model. *BMC Bioinformatics*, 19:161. 15

Zhang, S.-B. and Lai, J.-H. (2014). Semantic similarity measurement between gene ontology terms based on exclusively inherited shared information. *Gene*, 558. 16

Zhao, C. and Wang, Z. (2018). Gogo: An improved algorithm to measure the semantic similarity between gene ontology terms. *Scientific Reports*, 8. 16

Zhong, X., Kaalia, R., and Rajapakse, J. (2019). GO2Vec: transforming GO terms and proteins to vector representations via graph embeddings. *BMC Genomics*, 20. 18